

# Detecting Gradual Changes in Locally Stationary Processes

Michael Vogt

University of Konstanz

Holger Dette

Ruhr-Universität Bochum

March 18, 2014

## Abstract

In a wide range of applications, the stochastic properties of the observed time series change over time. The changes often occur gradually rather than abruptly: the properties are (approximately) constant for some time and then slowly start to change. In such situations, it is frequently of interest to locate the time point where the properties start to vary. In contrast to the analysis of abrupt changes, methods for detecting smooth or gradual change points are less developed and often require strong parametric assumptions. In this paper, we develop a fully nonparametric method to estimate a smooth change point in a locally stationary framework. We set up a general procedure which allows to deal with a wide variety of stochastic properties including the mean, (auto)covariances and higher-order moments. The theoretical part of the paper establishes the convergence rate of the new estimator. In addition, we examine its finite sample performance by means of a simulation study and illustrate the methodology by applications to temperature and financial return data.

**Key words:** Local stationarity; empirical processes; measure of time-variation, gradual changes.

**AMS 2010 subject classifications:** 62G05, 62G20, 62M10.

## 1 Introduction

In many applications, the stochastic properties of the observed time series such as the mean, the variance or the distribution change over time. In the classical structural break setting, the changes are abrupt: the stochastic properties are constant for some time and then suddenly jump to another value. In a number of situations, however, the changes occur gradually rather than abruptly: the properties are (approximately) constant for a while and then reach a time point where they slowly start to change. We refer to this time point as a smooth or gradual change point in what follows.

Locating a smooth change point is important in a wide range of applications. As a first example, consider the monthly temperature anomalies (temperature deviations from a reference value) of the northern hemisphere from 1850 to 2013 which are displayed in the left-hand panel of Figure 1. Global mean temperature records over the last 150 years suggest that there has been a significant upward trend in the temperature [see Bloomfield (1992) and

Hansen et al. (2002) among others]. This upward trend which is commonly termed “global warming” is also visible in the time series of Figure 1. Inspecting the plot more closely, the mean temperature appears to be fairly constant at the beginning of the sample and then starts to gradually increase. An important issue is to detect the advent of “global warming”, that is, the time point where the mean of the time series starts to trend upwards.

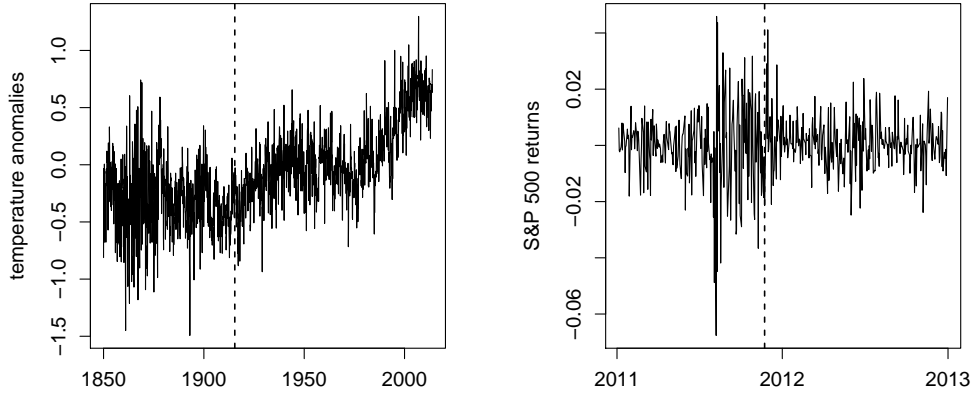


Figure 1: The left-hand panel shows the monthly temperature anomalies of the northern hemisphere from 1850 to 2013 measured in  $^{\circ}\text{C}$ . The right-hand panel depicts the daily returns of the S&P 500 index from the beginning of 2011 to the end of 2012. The vertical lines in the figures indicate the gradual change points estimated by the method developed in this paper.

A second example can be found in the right-hand panel of Figure 1 which shows the daily returns of the S&P 500 stock index from the beginning of 2011 to the end of 2012. Inspecting the data, it is apparent that the volatility level changes over time. Moreover, the plot suggests that the volatility is roughly constant in 2012 but gradually increases before that. Denoting the present time point by  $T$ , practitioners are often interested in identifying the time interval  $[t_0, T]$  where the volatility level is more or less constant. Put differently, they are interested in localizing the time point  $t_0$  prior to which the volatility starts to substantially vary over time. Once the time point  $t_0$  has been identified, it is common practice in volatility forecasting to fit a model to the data in the time span  $[t_0, T]$  [see e.g. Chen et al. (2010)].

Further examples can be found in a variety of different areas. In the analysis of EEG data, for instance, it is of interest to locate the time point where an epileptic seizure occurs. The onset of a seizure arguably coincides with a change in the autocovariance structure of the EEG data. The aim is thus to estimate the time point where the autocovariance structure starts to vary. Another example comes from economics and concerns the Great Moderation, that is, the reduction of the volatility level of business cycle fluctuations in the middle of the 1980s. This moderation is clearly visible in the time series of U.S. GDP data: The volatility level of the data is roughly stable until the mid 1980s and then starts to reduce. Here, it is of interest to pin down the time point where the moderation begins.

In most applications, there is not much known about the way in which the stochastic properties of interest evolve over time. For instance, it is not clear at all what is the functional form of the warming trend in our first example. There is no reason why it should have a particular parametric structure. Similarly, there is no economic theory suggesting that the increase of

the volatility level before 2012 in our second example should have a specific parametric form. It is thus important to have flexible nonparametric methods at hand which allow to locate a smooth change point without imposing strong parametric restrictions on the time-varying behaviour of the stochastic properties under consideration.

The main goal of this paper is to develop such a method. More precisely, we tackle the following estimation problem: Suppose we observe a sample of data  $\{X_{t,T} : t = 1, \dots, T\}$  and are interested in a stochastic feature such as the mean  $\mathbb{E}[X_{t,T}]$  or the variance  $\text{Var}(X_{t,T})$ . Moreover, assume that the feature is time-invariant on the time span  $\{1, \dots, t_0\}$ , or equivalently, on the rescaled time interval  $[0, u_0]$  with  $u_0 = t_0/T$  and then starts to gradually vary over time. Our aim is to estimate the rescaled time point  $u_0$ . We do not impose any parametric restrictions on the time-varying behaviour of the feature of interest after  $u_0$ . In this sense, our model setting is completely nonparametric. Moreover, rather than restricting attention to a specific stochastic property, we set up a general procedure which allows to deal with a wide variety of features including the mean, (auto)covariances and higher-order moments of the time series at hand. We tackle the problem of estimating  $u_0$  within a locally stationary framework which is well suited to model gradual changes and is formally introduced in Section 2.

The nonparametric nature of our estimation problem sharply distinguishes it from standard change point problems and requires new methodology. The literature commonly imposes strong parametric restrictions on the time-varying behaviour of the stochastic properties at hand. In the vast majority of papers, the changes are abrupt, that is, the properties are assumed to be constant over time apart from some occasional structural breaks. The detection of sudden structural breaks has a long history originating from statistical inference in quality control [see for example Page (1954, 1955) for some early references]. Since its introduction many authors have worked on this problem [see Chow (1960), Brown et al. (1975) or Krämer et al. (1988), among others]. Most of the literature investigates the issue of detecting breaks in the mean or the variance of a time series [see Horváth et al. (1999) or Aue et al. (2009)], the parameters of regression models [see Andrews (1993) or Bai and Perron (1998)] or the second order characteristics of a time series [see Berkes et al. (2009), Wied et al. (2012) or Davis et al. (2006)]. An extensive list of references on the localization of abrupt changes can be found in Jandhyala et al. (2014).

The literature on detecting gradual changes is much more scarce than that on abrupt changes. Most references consider location models of a very simple parametric form. For example, several authors investigate broken line regression models with independent normally distributed errors [see for example Hinkley (1970) or Siegmund and Zhang (1994)] and the performance of control charts under a gradual change in the mean [see Bissell (1984b,a) or Gan (1991, 1992) among others]. Other work considers estimators and tests in models where the linear drift has been replaced by some smooth parametric function (such as a polynomial) and the errors are assumed to be independent identically distributed but not necessarily normal [see Hušková (1999), Hušková and Steinebach (2002) and also Aue and Steinebach (2002) for a generalization to the dependent case].

More recently, there has been some work on the problem of detecting smooth change points

in some simple nonparametric settings. Most authors consider the location model  $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$  with zero mean i.i.d. errors  $\varepsilon_t$ . Indeed, in many cases, the errors are even assumed to be Gaussian. Suppose that the mean function  $\mu$  is constant on the interval  $[0, u_0]$ , i.e.,  $\mu(u) = \bar{\mu}$  for  $u \leq u_0$  and then starts to smoothly vary over time. Under appropriate smoothness conditions,  $u_0$  can be regarded as a break point in the  $k$ -th derivative of  $\mu$ . It can thus be estimated by methods to detect a break point in a higher-order derivative of a nonparametric function [see Müller (1992) for an early reference and e.g. Raimondo (1998) and Goldenshluger et al. (2006) who derive minimax rates in the model with Gaussian errors]. Mallik et al. (2011, 2013) propose an alternative  $p$ -value based approach to estimate  $u_0$  when  $\mu$  is a smooth nonparametric function that is restricted to take values larger than  $\bar{\mu}$  at time points  $u > u_0$ , that is,  $\mu(u) > \bar{\mu}$  for  $u > u_0$ . Finally, Mercurio and Spokoiny (2004) study sequential testing procedures for change point detection in some simple nonparametric volatility models. All these methods are tailored to a very specific model setting and often rely on strong distributional assumptions. Our procedure in contrast is very general in nature and can be applied to a wide variety of settings. Moreover, it does not rely on any distributional restrictions. In the location model  $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$ , for instance, we do not even require the errors to be independent or stationary. In fact, we are able to estimate  $u_0$  as long as the errors are locally stationary.

In Section 4, we introduce our estimator of the time point  $u_0$  which is based on a refinement of the CUSUM principle. To construct it, we proceed in two steps. In the first, we set up a function  $\mathcal{D} : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ , where  $\mathcal{D}(u)$  measures the amount of time-variation in the stochastic feature of interest within the interval  $[0, u]$ . By construction,  $\mathcal{D}(u) = 0$  if there is no time-variation on the interval  $[0, u]$  and  $\mathcal{D}(u) > 0$  if there is some time-variation involved. Since  $\mathcal{D}$  is not observed, we replace it by an estimator  $\hat{\mathcal{D}}_T$ . Section 3 gives a detailed account of how to construct the measure of time-variation  $\mathcal{D}$  and its estimator  $\hat{\mathcal{D}}_T$ . The time point  $u_0$  can now be characterized as the point where the measure  $\mathcal{D}$  starts to deviate from zero. This characterization is used in the second step to come up with an estimator of  $u_0$ . Section 4 describes in detail how to set up this estimator.

In Section 5, we examine the asymptotic properties of our approach. In particular, we show that the proposed estimator is consistent and derive its convergence rate. As we will see, the rate depends on the degree of smoothness of the stochastic feature of interest at  $u_0$ . This reflects the fact that it becomes harder to locate the time point  $u_0$  when the feature varies more slowly and smoothly around this point. Our method depends on a tuning parameter with a specific statistical interpretation. In particular, it is similar in nature to a critical value in a testing procedure and can be chosen to keep a pre-specified probability of underestimating the point  $u_0$ . We derive a data driven choice of the tuning parameter with good theoretical and practical properties in Sections 5.4 and 6. The first and second part of Section 7 investigate the small sample performance of our method by means of a simulation study and compare it with competing methods for the location model  $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$ . Additional simulations can be found in the Supplementary Material to the paper. Finally, in the third part of Section 7, we apply our method to the two data sets from Figure 1. Specifically, we use our procedure to estimate the advent of “global warming” and the time

point prior to which the volatility level of the S&P 500 returns strongly varies over time.

## 2 Model Setting

Throughout the paper, we assume that the sample of observations  $\{X_{t,T} : t = 1, \dots, T\}$  comes from a locally stationary process of  $d$ -dimensional variables  $X_{t,T}$ . Specifically, we work with the following concept of local stationarity, which was introduced in Vogt (2012).

**Definition 2.1.** *The array  $\{X_{t,T} : t = 1, \dots, T\}_{T=1}^\infty$  is called a locally stationary process if for each rescaled time point  $u \in [0, 1]$ , there exists a strictly stationary process  $\{X_t(u) : t \in \mathbb{Z}\}$  with the property that*

$$\|X_{t,T} - X_t(u)\| \leq \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right) U_{t,T}(u) \quad a.s.$$

Here,  $\|\cdot\|$  denotes a norm on  $\mathbb{R}^d$  and  $\{U_{t,T}(u) : t = 1, \dots, T\}_{T=1}^\infty$  is an array of positive random variables whose  $\rho$ -th moment is uniformly bounded for some  $\rho > 0$ , that is,  $\mathbb{E}[U_{t,T}^\rho(u)] \leq C < \infty$  for some fixed constant  $C$ .

Local stationarity was initially defined in terms of a time-varying spectral representation in Dahlhaus (1997). Our definition of local stationarity is similar to those in Dahlhaus and Subba Rao (2006) and Koo and Linton (2012) for example. The intuitive idea behind these definitions is that a process is locally stationary if it behaves approximately stationary locally in time, i.e., over short time periods. This idea is turned into a rigorous concept by requiring that locally around each rescaled time point  $u$ , the process  $\{X_{t,T}\}$  can be approximated by a stationary process  $\{X_t(u)\}$  in a stochastic sense.

There is a wide range of time series processes which are locally stationary in the sense of Definition 2.1. In particular, many processes with time-varying parameters can be locally approximated by a stationary process provided that the parameters are smoothly changing over time. This is fairly straightforward to show for linear models like time-varying MA or AR processes. However, it may also be verified for more complicated models like time-varying GARCH processes [see Dahlhaus and Subba Rao (2006) or Subba Rao (2006)].

The definition of local stationarity relies on rescaling time to the unit interval. The main reason for doing so is to obtain a reasonable asymptotic theory. Rescaling the time argument is also common in the investigation of change points. While a completely specified parametric model as considered in Hinkley (1970) or Siegmund and Zhang (1994) does not need this technique, more general approaches are usually based on rescaling arguments [see Hušková (1999) or Aue and Steinebach (2002) among others].

Let  $\lambda_{t,T}$  be some time-varying feature of the locally stationary process  $\{X_{t,T}\}$  such as the mean  $\mathbb{E}[X_{t,T}]$  or the variance  $\text{Var}(X_{t,T})$ . Generally speaking, we allow for any feature  $\lambda_{t,T}$  which fulfills the following property:

$(P_\lambda)$   $\lambda_{t,T}$  is uniquely determined by the set of moments  $\{\mathbb{E}[f(X_{t,T})] : f \in \mathcal{F}\}$ , where  $\mathcal{F}$  is a family of measurable functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .

Note that  $(P_\lambda)$  is a fairly weak condition which is satisfied by a wide range of stochastic features. Indeed, it essentially allows us to deal with any feature that can be expressed in terms of a set of moments. We illustrate the property  $(P_\lambda)$  by some examples:

**Example I.** Let  $\lambda_{t,T}$  be the mean  $\mu_{t,T} = \mathbb{E}[X_{t,T}]$  of a univariate locally stationary process  $\{X_{t,T}\}$ . Then the corresponding family of functions is simply  $\mathcal{F} = \{\text{id}\}$ , since the mean  $\mu_{t,T}$  can be written as  $\mathbb{E}[\text{id}(X_{t,T})]$ .

**Example II.** Let  $\lambda_{t,T}$  be the vector of the first  $p$  autocovariances of a univariate locally stationary process  $\{Y_{t,T}\}$  whose elements  $Y_{t,T}$  are centred for simplicity. Specifically, define  $\gamma_{\ell,t,T} = \text{Cov}(Y_{t,T}, Y_{t-\ell,T})$  to be the  $\ell$ -th order autocovariance and set  $\lambda_{t,T} = (\gamma_{0,t,T}, \dots, \gamma_{p,t,T})^\top$ . To handle this case, we regard the data as coming from the  $(p+1)$ -dimensional process  $\{X_{t,T}\}$  with  $X_{t,T} = (Y_{t,T}, Y_{t-1,T}, \dots, Y_{t-p,T})^\top$ . We now define functions  $f_\ell : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$  for  $0 \leq \ell \leq p$  by  $f_\ell(x) = x_0 x_\ell$ , where  $x = (x_0, \dots, x_p)^\top$ . As  $\mathbb{E}[f_\ell(X_{t,T})] = \mathbb{E}[Y_{t,T} Y_{t-\ell,T}] = \gamma_{\ell,t,T}$ , we obtain that  $\mathcal{F} = \{f_0, \dots, f_p\}$  in this setting.

**Example III.** Consider a  $d$ -dimensional locally stationary process  $\{X_{t,T}\}$  whose elements  $X_{t,T} = (X_{t,T,1}, \dots, X_{t,T,d})^\top$  are again centred for simplicity. Let  $\lambda_{t,T}$  be the vector of covariances  $\nu_{t,T}^{(i,j)} = \text{Cov}(X_{t,T,i}, X_{t,T,j})$ , that is,  $\lambda_{t,T} = (\nu_{t,T}^{(i,j)})_{1 \leq i \leq j \leq d}$ . Analogously as in the previous example,  $\mathcal{F} = \{f_{ij} : 1 \leq i \leq j \leq d\}$  with  $f_{ij}(x) = x_i x_j$ .

We next define  $\lambda(u)$  to be the stochastic feature of the approximating process  $\{X_t(u)\}$  which corresponds to  $\lambda_{t,T}$ . This means that  $\lambda(u)$  is fully characterized by the set of moments  $\{\mathbb{E}[f(X_t(u))] : f \in \mathcal{F}\}$ . Throughout the paper, we assume that

$$\sup_{f \in \mathcal{F}} |\mathbb{E}[f(X_{t,T})] - \mathbb{E}[f(X_t(u))]| \leq C \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right), \quad (2.1)$$

which is implied by the high-order condition (C4) in Subsection 5.1. In a wide range of cases, the inequality (2.1) boils down to mild moment conditions on the random variables  $X_{t,T}$ ,  $X_t(u)$  and  $U_{t,T}(u)$ . This in particular holds true in Examples I–III as discussed in Subsection 5.1. The inequality (2.1) essentially says that  $\lambda_{t,T}$  and  $\lambda(u)$  are close to each other locally in time. In the time-varying mean setting from Example I, it can be expressed as

$$|\mu_{t,T} - \mu(u)| \leq C \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right)$$

with  $\mu(u)$  being the mean of  $X_t(u)$ . In Example II, it is equivalent to the statement

$$\|(\gamma_{0,t,T}, \dots, \gamma_{p,t,T})^\top - (\gamma_0(u), \dots, \gamma_p(u))^\top\| \leq C \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right),$$

where  $\gamma_\ell(u) = \text{Cov}(Y_t(u), Y_{t-\ell}(u))$  and  $\|\cdot\|$  is some norm on  $\mathbb{R}^{p+1}$ . Similarly, in Example III, it says that

$$\|(\nu_{t,T}^{(i,j)})_{i,j=1,\dots,d} - (\nu^{(i,j)}(u))_{i,j=1,\dots,d}\| \leq C \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right),$$

where  $\nu^{(i,j)}(u) = \text{Cov}(X_{t,i}(u), X_{t,j}(u))$ . Hence, if (2.1) holds true, the feature  $\lambda_{t,T}$  converges to  $\lambda(u)$  locally in time. In particular, time-variation in  $\lambda_{t,T}$  is asymptotically equivalent to

time-variation in  $\lambda(u)$ . To detect whether the stochastic feature  $\lambda_{t,T}$  of interest changes over time, we may thus check for variations in the approximating quantity  $\lambda(u)$ .

Our estimation problem can now be formulated as follows: Assume that  $\lambda(u)$  does not vary on the rescaled time interval  $[0, u_0]$  but is time-varying after  $u_0$ . Our aim is to estimate the time point  $u_0$  where  $\lambda(u)$  starts to change over time.

### 3 A Measure of Time-Variation

In this section, we set up a function  $\mathcal{D} : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$  which captures time-variations in the stochastic feature  $\lambda = \lambda(\cdot)$  of interest and explain how to estimate it. By construction, the function  $\mathcal{D}$  has the property

$$(P_{\mathcal{D}}) \quad \mathcal{D}(u) \begin{cases} = 0 & \text{if } \lambda \text{ does not vary on } [0, u] \\ > 0 & \text{if } \lambda \text{ varies on } [0, u] \end{cases}$$

and is called a measure of time-variation. In what follows, we describe how to set up such a measure for a generic stochastic feature that satisfies  $(P_{\lambda})$ .

Our construction is based on the following idea: By the property  $(P_{\lambda})$ , the feature  $\lambda(w)$  is fully characterized by the values  $\mathbb{E}[f(X_t(w))]$  with  $f$  running over all functions in the family  $\mathcal{F}$ . This implies that time-variation in  $\lambda(w)$  is equivalent to time-variation in the moments  $\mathbb{E}[f(X_t(w))]$  for some  $f \in \mathcal{F}$ . To detect changes in  $\lambda(w)$  over time, we may thus set up a function which captures time-variations in the quantities  $\mathbb{E}[f(X_t(w))]$  for any  $f \in \mathcal{F}$ . This idea underlies the following definition:

$$\mathcal{D}(u) = \sup_{f \in \mathcal{F}} \sup_{v \in [0, u]} |D(u, v, f)|, \quad (3.1)$$

where

$$D(u, v, f) = \int_0^v \mathbb{E}[f(X_t(w))]dw - \left(\frac{v}{u}\right) \int_0^u \mathbb{E}[f(X_t(w))]dw. \quad (3.2)$$

If the moment function  $\mathbb{E}[f(X_t(\cdot))]$  is constant on the interval  $[0, u]$ , then the average  $v^{-1} \int_0^v \mathbb{E}[f(X_t(w))]dw$  takes the same value at all points  $v \in [0, u]$ . From this, it immediately follows that  $D(u, v, f) = 0$  for any  $v \in [0, u]$ . Hence, if the function  $\mathbb{E}[f(X_t(\cdot))]$  is constant on  $[0, u]$  for any  $f \in \mathcal{F}$ , then the measure of time-variation satisfies  $\mathcal{D}(u) = 0$ . If  $\mathbb{E}[f(X_t(\cdot))]$  varies on  $[0, u]$  for some  $f$  in contrast, then the average  $v^{-1} \int_0^v \mathbb{E}[f(X_t(w))]dw$  varies on this time span as well. This is ensured by the fact that  $\mathbb{E}[f(X_t(\cdot))]$  is a Lipschitz continuous function of rescaled time, i.e.,  $|\mathbb{E}[f(X_t(w))] - \mathbb{E}[f(X_t(w'))]| \leq C|w - w'|$  for any  $w, w' \in [0, 1]$ , which is a direct consequence of (2.1). We thus obtain that  $D(u, v, f) > 0$  for some  $v \in [0, u]$ , which in turn yields that  $\mathcal{D}(u) > 0$ . As a result,  $\mathcal{D}$  satisfies  $(P_{\mathcal{D}})$ .

Since the feature  $\lambda$  is constant on  $[0, u_0]$  but varies after  $u_0$ , the property  $(P_{\mathcal{D}})$  immediately implies that  $\mathcal{D}(u) = 0$  for  $u \leq u_0$  and  $\mathcal{D}(u) > 0$  for  $u > u_0$ . The point  $u_0$  is thus characterized as the time point where the measure of time-variation starts to deviate from zero. Importantly, the measure  $\mathcal{D}$  does not have a jump at  $u_0$ , but smoothly deviates from zero at this point. Its degree of smoothness depends on how smoothly the moments  $\mathbb{E}[f(X_t(w))]$

vary over time, or put differently, on how smoothly the feature  $\lambda(w)$  varies over time. In particular, the smoother the time-variation in  $\lambda$ , the smoother the function  $\mathcal{D}$ .

In order to estimate the measure of time-variation, we proceed as follows: The integral  $\int_0^v \mathbb{E}[f(X_t(w))]dw$  can be regarded as an average of the moments  $\mathbb{E}[f(X_t(w))]$ , where all time points from 0 to  $v$  are taken into account. This suggests to estimate it by a sample average of the form  $T^{-1} \sum_{t=1}^{\lfloor vT \rfloor} f(X_{t,T})$ . Following this idea, an estimator of  $\mathcal{D}(u)$  is given by

$$\hat{\mathcal{D}}_T(u) = \sup_{f \in \mathcal{F}} \sup_{v \in [0, u]} |\hat{D}_T(u, v, f)|,$$

where we set

$$\hat{D}_T(u, v, f) = \frac{1}{T} \sum_{t=1}^{\lfloor vT \rfloor} f(X_{t,T}) - \left(\frac{v}{u}\right) \frac{1}{T} \sum_{t=1}^{\lfloor uT \rfloor} f(X_{t,T}).$$

The statistic  $\hat{\mathcal{D}}_T(u)$  is constructed by the CUSUM principle for the interval  $[0, u]$  and can be regarded as a generalization of classical CUSUM statistics to be found for example in Page (1954, 1955). The quantity  $\hat{D}_T$  compares cumulative sums of the variables  $f(X_{t,T})$  over different time spans  $[0, v]$  and  $[0, u]$ . By taking the supremum with respect to  $v \in [0, u]$ , we are able to detect gradual changes in the signal  $\mathbb{E}[f(X_t(\cdot))]$  on the interval  $[0, u]$ . The additional supremum over  $f$  makes sure that the signals corresponding to all functions  $f \in \mathcal{F}$  are taken into account.

## 4 Estimating the Gradual Change Point $u_0$

We now describe how to use our measure of time-variation to estimate the point  $u_0$ . Our estimation method is based on the observation that  $\sqrt{T}\mathcal{D}(u) = 0$  for  $u \leq u_0$  and  $\sqrt{T}\mathcal{D}(u) \rightarrow \infty$  for  $u > u_0$  as  $T \rightarrow \infty$ . The scaled estimator  $\sqrt{T}\hat{\mathcal{D}}_T(u)$  behaves in a similar way: As we will see later on,

$$\sqrt{T}\hat{\mathcal{D}}_T(u) \begin{cases} \xrightarrow{d} \mathcal{H}(u) & \text{for } u \leq u_0 \\ \xrightarrow{P} \infty & \text{for } u > u_0, \end{cases} \quad (4.1)$$

where  $\mathcal{H}(u)$  is a real-valued random variable. By (4.1),  $\sqrt{T}\hat{\mathcal{D}}_T(u)$  can be regarded as a statistic to test the hypothesis that the feature of interest  $\lambda$  is time-invariant on the interval  $[0, u]$ . Under the null of time-invariance, that is, as long as  $u \leq u_0$ , the statistic weakly converges to some limit distribution. Under the alternative, that is, at time points  $u > u_0$ , it diverges in probability to infinity. The main idea of the new estimation method is to exploit this dichotomous behaviour.

To construct our estimator of  $u_0$ , we proceed as follows: First of all, we define the quantity

$$\hat{r}_T(u) = 1(\sqrt{T}\hat{\mathcal{D}}_T(u) \leq \tau_T),$$

where  $\tau_T$  is a threshold level that slowly diverges to infinity. A data driven choice of  $\tau_T$  with good theoretical and practical properties is discussed in detail in Section 5.4. The random



variable  $\hat{r}_T(u)$  specifies the outcome of our test on time-invariance for the interval  $[0, u]$  given the critical value  $\tau_T$ : if the test accepts the null of time-invariance, then  $\hat{r}_T(u) = 1$ ; if it rejects the null, then  $\hat{r}_T(u) = 0$ . Under the null, the test statistic tends to take moderate values, suggesting that  $\hat{r}_T(u)$  should eventually become zero. Under the alternative, the statistic explodes, implying that  $\hat{r}_T(u)$  should finally take the value one. Formally speaking, one can show that

$$\hat{r}_T(u) \xrightarrow{P} \begin{cases} 1 & \text{for } u \leq u_0 \\ 0 & \text{for } u > u_0, \end{cases}$$

if  $\tau_T$  converges (slowly) to infinity. This suggests that  $\int_0^1 \hat{r}_T(u) du \approx u_0$  for large sample sizes. Hence, we may simply estimate  $u_0$  by aggregating the test outcomes  $\hat{r}_T(u)$ , that is,

$$\hat{u}_0(\tau_T) = \int_0^1 \hat{r}_T(u) du.$$

This estimator exploits the fact that the test outcome should be equal to one at time points  $u \leq u_0$  but equal to zero at  $u > u_0$ .

## 5 Asymptotic Properties

We now examine the asymptotic properties of the proposed estimation method. We first investigate the weak convergence behaviour of the statistic  $\hat{D}_T$  and then derive the convergence rate of the estimator  $\hat{u}_0(\tau_T)$ . Since the proofs are very technical and involved, they are deferred to the Appendix. To state the results, we let the symbol  $\ell_\infty(S)$  denote the space of bounded functions  $f : S \rightarrow \mathbb{R}$  endowed with the supremum norm and let  $\rightsquigarrow$  denote weak convergence. Moreover, to capture the amount of smoothness of the measure  $\mathcal{D}$  at the point  $u_0$ , we suppose that

$$\frac{\mathcal{D}(u)}{(u_0 - u)^\kappa} \rightarrow c_\kappa > 0 \quad \text{as } u \searrow u_0 \quad (5.1)$$

for some number  $\kappa > 0$  and a constant  $c_\kappa > 0$ . The larger  $\kappa$ , the more smoothly the measure  $\mathcal{D}$  deviates from zero at the point  $u_0$ .

### 5.1 Assumptions

Throughout the paper, we make the following assumptions:

- (C1) The process  $\{X_{t,T}\}$  is locally stationary in the sense of Definition 2.1.
- (C2) The process  $\{X_{t,T}\}$  is strongly mixing with mixing coefficients  $\alpha(k)$  satisfying  $\alpha(k) \leq Ca^k$  for some positive constants  $C$  and  $a < 1$ .
- (C3) Let  $p \geq 4$  be an even natural number and endow the set  $\mathcal{F}$  with some semimetric  $d_{\mathcal{F}}$ .  $(\mathcal{F}, d_{\mathcal{F}})$  is separable, compact and not too complex in the sense that its covering number  $\mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})$  satisfies the condition  $\int_0^1 \mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})^{1/p} dw < \infty$ . Moreover, the set  $\mathcal{F}$  has

an envelope  $F$  (i.e.  $|f| \leq F$  for all  $f \in \mathcal{F}$ ) which satisfies  $\mathbb{E}[F(X_{t,T})^{(1+\delta)p}] \leq C < \infty$  for some small  $\delta > 0$  and a fixed constant  $C$ . Finally, for any pair of functions  $f, f' \in \mathcal{F}$ ,

$$\mathbb{E}\left[\left|\frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f, f')}\right|^{(1+\delta)p}\right] \leq C < \infty.$$

(C4) For  $k = 1, 2$  and all  $f \in \mathcal{F}$ , it holds that  $\mathbb{E}[|f(X_{t,T}) - f(X_t(u))|^k] \leq C(|\frac{t}{T} - u| + \frac{1}{T})$  for some fixed constant  $C$ .

Condition (C2) stipulates that the array  $\{X_{t,T}\}$  is strongly mixing. A wide variety of locally stationary processes can be shown to be mixing under appropriate conditions; see for example Fryzlewicz and Subba Rao (2011) and Vogt (2012). To keep the structure of the proofs as clear as possible, we have assumed the mixing rates to decay exponentially fast. Alternatively, we could work with slower polynomial rates at the cost of a more involved notation in the proofs. Conditions (C3) and (C4) allow for a wide range of function families  $\mathcal{F}$  and are formulated in a very general way. For many choices of  $\mathcal{F}$ , they boil down to simple moment conditions on the variables  $X_{t,T}$ ,  $X_t(u)$  and  $U_{t,T}(u)$ . We illustrate this by means of Examples I–III. It is straightforward to show that in Example I, (C3) and (C4) are satisfied under the following set of moment conditions:

(A<sub>μ</sub>) Either (a)  $\mathbb{E}|X_{t,T}|^r \leq C$  for some  $r > 4$  and  $\mathbb{E}U_{t,T}^2(u) \leq C$  or (b)  $\mathbb{E}|X_{t,T}|^r \leq C$ ,  $\mathbb{E}|X_t(u)|^r \leq C$  and  $\mathbb{E}U_{t,T}^{r/(r-1)}(u) \leq C$  for some  $r > 4$  and a sufficiently large constant  $C$  that is independent of  $u, t$  and  $T$ .

Similarly, in Example II, they are implied by

(A<sub>γ</sub>)  $\mathbb{E}\|X_{t,T}\|^r \leq C$ ,  $\mathbb{E}\|X_t(u)\|^r \leq C$  and  $\mathbb{E}U_{t,T}^q(u) \leq C$  for some  $r > 8$  and  $q = \frac{r}{3}/(\frac{r}{3} - 1)$ , where  $C$  is a sufficiently large constant that is independent of  $u, t$  and  $T$ .

The moment conditions in Example III are fully analogous to those in Example II and thus not stated explicitly.

## 5.2 Weak convergence of the measure of time-variation

To start with, we investigate the asymptotic properties of the expression

$$\hat{H}_T(u, v, f) = \sqrt{T}(\hat{D}_T(u, v, f) - D(u, v, f)). \quad (5.2)$$

To do so, let  $\Delta = \{(u, v) \in [0, 1]^2 : v \leq u\}$  and equip the space  $\Delta \times \mathcal{F}$  with the natural semimetric  $|u - u'| + |v - v'| + d_{\mathcal{F}}(f, f')$ . In what follows, we regard  $\hat{H}_T$  as a process that takes values in  $\ell_{\infty}(\Delta \times \mathcal{F})$  and show that it weakly converges to a Gaussian process  $H$  with

the covariance structure

$$\begin{aligned} \text{Cov}(H(u, v, f), H(u', v', f')) &= \sum_{l=-\infty}^{\infty} \left\{ \frac{vv'}{uu'} \int_0^{\min\{u, u'\}} c_l(w) dw - \frac{v'}{u'} \int_0^{\min\{v, u'\}} c_l(w) dw \right. \\ &\quad \left. - \frac{v}{u} \int_0^{\min\{u, v'\}} c_l(w) dw + \int_0^{\min\{v, v'\}} c_l(w) dw \right\}, \end{aligned} \quad (5.3)$$

where  $c_l(w) = c_l(w, f, f') = \text{Cov}(f(X_0(w)), f'(X_l(w)))$ . The following theorem gives a precise description of the weak convergence of  $\hat{H}_T$ .

**Theorem 5.1.** *Let assumptions (C1)–(C4) be satisfied. Then*

$$\hat{H}_T = \sqrt{T}[\hat{D}_T - D] \rightsquigarrow H$$

as a process in  $\ell_\infty(\Delta \times \mathcal{F})$ , where  $\hat{D}_T$  and  $D$  are defined in Section 3 and  $H$  is a Gaussian process on  $\Delta \times \mathcal{F}$  with covariance kernel (5.3).

This theorem is the main stepping stone to derive the asymptotic properties of our estimator  $\hat{u}_0(\tau_T)$ . In addition, it is useful to examine the asymptotic behaviour of some processes related to  $\hat{H}_T$ : Analogously to  $\hat{D}_T(u)$ , we introduce the expression

$$\hat{\mathcal{H}}_T(u) = \sup_{f \in \mathcal{F}} \sup_{v \in [0, u]} |\hat{H}_T(u, v, f)|. \quad (5.4)$$

Moreover, we let

$$\hat{\mathbb{D}}_T(u) = \sup_{v \in [0, u]} \hat{D}_T(v) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u} |\hat{D}_T(v, w, f)| \quad (5.5)$$

together with

$$\hat{\mathbb{H}}_T(u) = \sup_{v \in [0, u]} \hat{\mathcal{H}}_T(v) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u} |\hat{H}_T(v, w, f)|. \quad (5.6)$$

The next result directly follows from Theorem 5.1 together with the continuous mapping theorem.

**Corollary 5.2.** *Let assumptions (C1)–(C4) be satisfied. Then*

$$\hat{\mathcal{H}}_T \rightsquigarrow \mathcal{H} \quad \text{and} \quad \hat{\mathbb{H}}_T \rightsquigarrow \mathbb{H}$$

as processes in  $\ell_\infty([0, 1])$ , where  $\mathcal{H}$  and  $\mathbb{H}$  are defined by  $\mathcal{H}(u) = \sup_{f \in \mathcal{F}, v \in [0, u]} |H(u, v, f)|$  and  $\mathbb{H}(u) = \sup_{f \in \mathcal{F}, 0 \leq w \leq v \leq u} |H(v, w, f)|$ , respectively.

### 5.3 Convergence of the estimator $\hat{u}_0(\tau_T)$

We now turn to the asymptotic behaviour of the estimator  $\hat{u}_0(\tau_T)$ . The next theorem shows that  $\hat{u}_0(\tau_T)$  consistently estimates  $u_0$  provided that the threshold level  $\tau_T$  diverges to infinity. Moreover, it specifies the convergence rate at which  $\hat{u}_0(\tau_T)$  approaches  $u_0$ .

**Theorem 5.3.** *Let assumptions (C1)–(C4) be satisfied and assume that  $\tau_T \rightarrow \infty$  with  $\tau_T/\sqrt{T} \rightarrow 0$ . Then*

$$\hat{u}_0(\tau_T) - u_0 = O_p(\gamma_T),$$

where  $\gamma_T = (\tau_T/\sqrt{T})^{1/\kappa}$  and  $\kappa$  is defined in (5.1).

The convergence rate of  $\hat{u}_0(\tau_T)$  can be seen to depend on the degree of smoothness  $\kappa$  of the measure  $\mathcal{D}$  at the point  $u_0$ . In particular, the smoother  $\mathcal{D}$ , the slower the convergence rate. Since the smoothness of  $\mathcal{D}$  mirrors that of the stochastic feature  $\lambda$ , we can equivalently say: the smoother the feature  $\lambda$  varies around  $u_0$ , the slower the rate of our estimator gets. This reflects the intuition that it becomes harder to precisely localize the point  $u_0$  when  $\lambda$  varies more smoothly and gradually around this point. The rate  $\gamma_T$  also depends on the threshold parameter  $\tau_T$ . Specifically, the slower  $\tau_T$  diverges to infinity, the faster the rate  $\gamma_T$  goes to zero. Hence, from a theoretical point of view,  $\tau_T$  should be chosen to diverge as slowly as possible to speed up the convergence rate of the estimator.

## 5.4 Choice of the threshold level $\tau_T$

We next discuss how to choose the threshold  $\tau_T$  to obtain an estimator of  $u_0$  with good theoretical properties. To state the results, we let  $q_\alpha(u)$  be the  $(1 - \alpha)$ -quantile of the limiting variable  $\mathbb{H}(u)$  and assume throughout that this quantile is known for any time point  $u$ . In practice, it is indeed unknown and has to be approximated. We show how to achieve this in Section 6 where we discuss the implementation of our method. Our choice of the threshold  $\tau_T$  proceeds in two steps. In the first, we describe a rough choice of  $\tau_T$  which leads to a preliminary estimator of  $u_0$ . In the second, we use this preliminary estimator to come up with a refined choice of  $\tau_T$  which in turn yields a better estimator of  $u_0$ .

**Preliminary choice of  $\tau_T$ .** To convey the idea behind the choice of  $\tau_T$ , let us first assume that  $\tau_T$  does not depend on the sample size, i.e.,  $\tau_T = \tau$  for all  $T$ . A first crude choice of  $\tau$  can be obtained by arguing in a similar way as in classical change point detection problems: Consider the situation that the stochastic feature of interest is time-invariant on  $[0, 1]$ , i.e., there is no change point  $u_0 < 1$ . In this situation, we would like to control the probability of false detection of a change point. Specifically, we aim to choose  $\tau$  such that this probability is smaller than some pre-specified level  $\alpha$ , that is,

$$\mathbb{P}(\hat{u}_0(\tau) < 1) \leq \alpha,$$

when there is no change point  $u_0 < 1$ . To achieve this, we write

$$\mathbb{P}(\hat{u}_0(\tau) < 1) \leq \mathbb{P}(\sqrt{T}\hat{\mathcal{D}}_T(u) > \tau \text{ for some } u \in [0, 1]) = \mathbb{P}(\sqrt{T}\hat{\mathbb{D}}_T(1) > \tau).$$

Corollary 5.2 shows that  $\sqrt{T}\hat{\mathbb{D}}_T(u)$  weakly converges to the limiting variable  $\mathbb{H}(u)$  at each point  $u \leq u_0$ . In particular, when there is no change point  $u_0 < 1$ , it holds that  $\sqrt{T}\hat{\mathbb{D}}_T(1) \xrightarrow{d} \mathbb{H}(1)$ . We now set  $\tau$  to be the  $(1 - \alpha)$ -quantile  $q_\alpha(1)$  of  $\mathbb{H}(1)$ . Writing  $\tau_\alpha^\circ = q_\alpha(1)$ , we obtain

that

$$\mathbb{P}(\hat{u}_0(\tau_\alpha^\circ) < 1) \leq \alpha + o(1),$$

when there is no change point  $u_0 < 1$ . We are thus able to asymptotically control the probability of false detection by choosing  $\tau = \tau_\alpha^\circ$ . However, this choice does not yield a consistent estimator of  $u_0$ . To ensure consistency, we have to make sure that the threshold  $\tau_T$  diverges to infinity. To achieve this, we let the level  $\alpha_T$  depend on the sample size  $T$ . In particular, we let it slowly converge to zero and set  $\tau_T = \tau_{\alpha_T}^\circ$ .

**Refined choice of  $\tau_T$ .** As in classical change point problems, the choice  $\tau = \tau_\alpha^\circ$  is fairly conservative. In particular, the resulting estimator tends to strongly overestimate the time point  $u_0$ . In what follows, we refine the choice of  $\tau$  to get a more precise estimator of  $u_0$ . Rather than controlling the false detection rate, we would like to control the probability of underestimating  $u_0$ , i.e., of falsely detecting a change point before it occurs. Technically speaking, we aim to choose  $\tau$  such that

$$\mathbb{P}(\hat{u}_0(\tau) < u_0) \leq \alpha$$

for some given level  $\alpha$ . Similarly as above, it holds that

$$\mathbb{P}(\hat{u}_0(\tau) < u_0) \leq \mathbb{P}(\sqrt{T}\hat{\mathcal{D}}_T(u) > \tau \text{ for some } u \in [0, u_0]) = \mathbb{P}(\sqrt{T}\hat{\mathbb{D}}_T(u_0) > \tau).$$

By Corollary 5.2, we know that  $\sqrt{T}\hat{\mathbb{D}}_T(u_0) \xrightarrow{d} \mathbb{H}(u_0)$ . Setting  $\tau$  to equal the  $(1-\alpha)$ -quantile  $q_\alpha(u_0)$  of the limiting variable  $\mathbb{H}(u_0)$  and using the notation  $\tau_\alpha = q_\alpha(u_0)$ , we are able to derive the following result.

**Theorem 5.4.** *Let assumptions (C1)–(C4) be satisfied. Then*

$$\mathbb{P}(\hat{u}_0(\tau_\alpha) < u_0) \leq \alpha + o(1) \tag{5.7}$$

and for any constant  $C > 0$ ,

$$\mathbb{P}(\hat{u}_0(\tau_\alpha) > u_0 + C\gamma_T) = o(1), \tag{5.8}$$

where  $\gamma_T$  is defined in Theorem 5.3.

Hence, the estimator  $\hat{u}_0(\tau_\alpha)$  has the following properties: According to (5.7), the probability of underestimating  $u_0$  is asymptotically bounded by  $\alpha$ . Moreover, the probability of overestimating  $u_0$  by more than  $C\gamma_T$  is asymptotically negligible by (5.8). Thus,  $\hat{u}_0(\tau_\alpha)$  controls the error of underestimating  $u_0$  while being consistent when it comes to overestimation.

Of course, we cannot take the choice  $\tau = \tau_\alpha$  at face value since the quantile  $\tau_\alpha = q_\alpha(u_0)$  depends on the unknown location  $u_0$ . Nevertheless, we can estimate this quantile by  $\hat{\tau}_\alpha = q_\alpha(\hat{u}_0(\tau_T^\circ))$ , where  $\hat{u}_0(\tau_T^\circ)$  is a consistent pilot estimator of  $u_0$ . In particular, we may set  $\tau_T^\circ = \tau_{\alpha_T}^\circ$  and use  $\hat{u}_0(\tau_{\alpha_T}^\circ)$  as a pilot estimate. It is fairly straightforward to see that the statements of Theorem 5.4 still hold true when  $\tau_\alpha$  is replaced by  $\hat{\tau}_\alpha$ :

**Corollary 5.5.** *Let assumptions (C1)–(C4) be satisfied. Then*

$$\mathbb{P}(\hat{u}_0(\hat{\tau}_\alpha) < u_0) \leq \alpha + o(1) \quad (5.9)$$

and for any  $C > 0$ ,

$$\mathbb{P}(\hat{u}_0(\hat{\tau}_\alpha) > u_0 + C\gamma_T) = o(1). \quad (5.10)$$

As in the previous subsection, we suggest to set  $\tau_T = \hat{\tau}_{\alpha_T}$  with  $\alpha_T$  gradually converging to zero to obtain a consistent estimator of  $u_0$ .

## 6 Implementation

To implement our estimation method in practice, we proceed as follows:

*Step 1.* Fix a probability level  $\alpha$  and estimate the threshold parameter  $\tau_\alpha$ .

- (i) Approximate the quantiles  $q_\alpha(u)$  by  $\hat{q}_\alpha(u)$  as described below.
- (ii) Compute the preliminary estimator  $\hat{u}_0^\circ = \hat{u}_0(\hat{\tau}_\alpha^\circ)$ , where  $\hat{\tau}_\alpha^\circ = \hat{q}_\alpha(1)$ .
- (iii) Estimate  $\tau_\alpha$  by  $\hat{\tau}_\alpha = \hat{q}_\alpha(\hat{u}_0^\circ)$ .

*Step 2.* Estimate  $u_0$  by  $\hat{u}_0(\hat{\tau}_\alpha)$ .

Generally speaking, the quantiles  $q_\alpha(u)$  can be approximated as follows: By definition,

$$\mathbb{H}(u) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u} |H(v, w, f)|$$

is the supremum of the Gaussian process  $H$  whose covariance structure is given in (5.3). Inspecting the formula (5.3), the only unknown expressions occurring in it are of the form

$$\sigma^2(u, f, f') = \sum_{l=-\infty}^{\infty} \Gamma_l(u, f, f'),$$

where  $\Gamma_l(u, f, f') = \int_0^u c_l(w, f, f') dw$  and  $c_l(w, f, f') = \text{Cov}(f(X_0(w)), f'(X_l(w)))$ . These quantities are essentially average long-term covariances of the processes  $\{f(X_t(w))\}$  and  $\{f'(X_t(w))\}$  on the interval  $[0, u]$ , which can be estimated by methods for long-run covariance estimation. Specifically, we can employ HAC type estimation procedures as discussed in Andrews (1991) or de Jong and Davidson (2000) among others and work with an estimator of the form

$$\hat{\sigma}^2(u, f, f') = \sum_{l=-\infty}^{\infty} K\left(\frac{l}{b(T)}\right) \hat{\Gamma}_l(u, f, f'). \quad (6.1)$$

where  $K$  is a kernel of Bartlett or flat-top type and  $b = b(T)$  is the bandwidth. Moreover,

$$\hat{\Gamma}_l(u, f, f') = \frac{1}{T} \sum_{t=1}^{\lfloor uT \rfloor} \hat{Z}_{t,T}(f) \hat{Z}_{t-l,T}(f'),$$

where  $\hat{Z}_{t,T}(f) = f(X_{t,T}) - \hat{\mathbb{E}}[f(X_{t,T})]$  and  $\hat{\mathbb{E}}[f(X_{t,T})]$  is an estimator of  $\mathbb{E}[f(X_{t,T})]$ . We may for example use a Nadaraya-Watson estimate

$$\hat{\mathbb{E}}[f(X_{t,T})] = \frac{1}{T} \sum_{s=1}^T K_h\left(\frac{t}{T} - \frac{s}{T}\right) f(X_{s,T})$$

with  $K$  being a kernel function and  $K_h(x) = h^{-1}K(x/h)$ . Alternatively, a local linear or more generally a local polynomial smoother may be employed. Once we have calculated the estimator  $\hat{\sigma}^2(u, f, f')$ , we can compute the covariance function (5.3) and simulate observations from the Gaussian process with the estimated covariance structure. This in turn allows us to simulate the quantiles  $q_\alpha(u)$ .

Our implementation strategy works well in practice as we will demonstrate in the empirical part of the paper. When the class of functions  $\mathcal{F}$  is large, it becomes computationally more burdensome to simulate the quantiles  $q_\alpha(u)$ . In most applications, however, the class of functions is fairly small. Moreover, in a number of cases, it is possible to simplify the implementation by exploiting the special structure of the model at hand. To illustrate this, we revisit the simple time-varying mean setting from Example I. In particular, we consider the model

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \varepsilon_{t,T}, \quad (6.2)$$

where the feature  $\lambda_{t,T}$  of interest is given by the mean function  $\mu(\frac{t}{T})$ . Recalling that  $\mathcal{F} = \{\text{id}\}$  in this case, the covariance structure (5.3) depends on the expressions  $\sigma^2(u) = \sum_{l=-\infty}^{\infty} \int_0^u c_l(w) dw$ , where  $c_l(w) = \mathbb{E}[\varepsilon_0(w)\varepsilon_l(w)]$  and  $\{\varepsilon_t(w)\}$  is the stationary approximating process of  $\{\varepsilon_{t,T}\}$  at the time point  $w$ . If the error process is stationary, we even obtain that  $\sigma^2(u) = u\sigma^2$  for all  $u$ , where  $\sigma^2 = \sum_{l=-\infty}^{\infty} \mathbb{E}[\varepsilon_0\varepsilon_l]$  is the long-run variance of the error terms. The latter can be estimated by standard methods. Denoting its estimator by  $\hat{\sigma}^2$ , we can set up our method in terms of the scaled statistic  $\hat{\mathcal{D}}_T^{\text{sc}}(u) = \hat{\mathcal{D}}_T(u)/\hat{\sigma}$ . Defining the expressions  $\hat{D}_T^{\text{sc}}(u)$ ,  $\hat{H}_T^{\text{sc}}(u)$  etc. in an analogous way, we obtain that  $\hat{H}_T^{\text{sc}} \rightsquigarrow H^{\text{sc}}$ , where the Gaussian process  $H^{\text{sc}}$  has the covariance structure

$$\text{Cov}(H^{\text{sc}}(u, v), H^{\text{sc}}(u', v')) = \frac{vv'}{uu'} \min\{u, u'\} - \frac{v'}{u'} \min\{v, u'\} - \frac{v}{u} \min\{u, v'\} + \min\{v, v'\}.$$

Importantly, this formula does not involve any unknown quantities, which in turn means that the quantiles  $q_\alpha^{\text{sc}}(u)$  of  $\mathbb{H}^{\text{sc}}(u)$  are completely known (neglecting the simulation error). Consequently, in this setting, which is often of interest in statistical practice, the method is particularly easy to implement.

## 7 Finite Sample Properties

### 7.1 Simulations

In this and the following subsection, we examine the small sample performance of our estimation method in a Monte-Carlo experiment. We first investigate two simulation setups

which are motivated by the applications in the introduction: a time-varying mean model and a volatility model together with a multivariate extension of it. Due to space constraints, the results on the volatility models are presented in the Supplementary Material. Here, we examine the setting

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \varepsilon_t \quad (7.1)$$

with two different mean functions  $\mu_1$  and  $\mu_2$ . The residuals  $\varepsilon_t$  are assumed to follow the AR(1) process  $\varepsilon_t = 0.25\varepsilon_{t-1} + \eta_t$ , where the innovations  $\eta_t$  are i.i.d. normal with zero mean and standard deviation 0.5. The mean functions are given by

$$\mu_1(u) = 1(u > 0.5) \quad (7.2)$$

$$\mu_2(u) = \{10(u - 0.5)\} \cdot 1(0.5 < u < 0.6) + 1(u > 0.6). \quad (7.3)$$

Both functions are equal to zero on the interval  $[0, 0.5]$  and then start to vary over time. Hence,  $u_0 = 0.5$  in both cases. The function  $\mu_1$  is a step function which allows to investigate how our method works in the presence of abrupt changes. The function  $\mu_2$  in contrast varies smoothly over time. In particular, it starts to linearly deviate from zero at the point  $u_0 = 0.5$  until it reaches a value of one and then remains constant.

To estimate the point  $u_0$ , we use the implementation strategy from Section 6 and denote the resulting estimator by  $\hat{u}_0$ . We set the parameter  $\alpha$  to equal 0.1 in all our simulations, meaning that the probability of underestimating  $u_0$  is approximately 10%. Moreover, as described at the end of Section 6, we normalize the statistic  $\hat{\mathcal{D}}_T(u)$  by an estimate of the long-run error variance  $\sigma^2 = \sum_{l=-\infty}^{\infty} \mathbb{E}[\varepsilon_0 \varepsilon_l]$ . To do so, we first approximate the residuals  $\varepsilon_t$  by  $\hat{\varepsilon}_t = X_{t,T} - \hat{\mu}_h(\frac{t}{T})$ , where  $\hat{\mu}_h$  is a Nadaraya-Watson estimator of  $\mu$ , and then apply a HAC estimator with a Bartlett kernel to the estimated residuals. Here, we set  $h = 0.2$  and choose the bandwidth of the HAC estimator to equal 10, i.e., we take into account the first ten autocovariances. As a robustness check, we have repeated the simulations for different bandwidth parameters. As this yields very similar results, we have not reported them here.



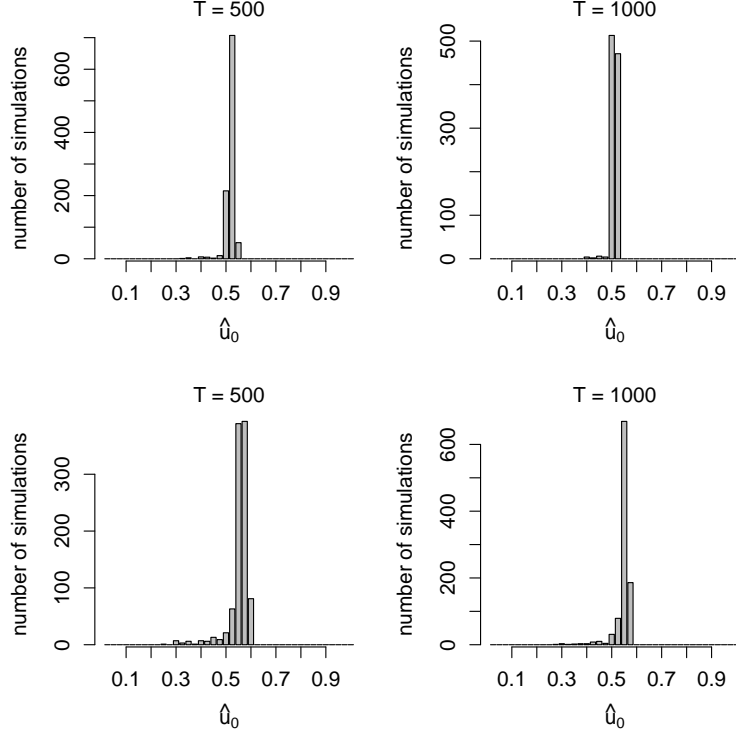


Figure 2: Simulation results produced by our method in model (7.1). Upper panel: results for the mean function  $\mu_1$  defined in (7.2). Lower panel: results for the mean function  $\mu_2$  defined in (7.3).

For each model setting, we produce  $N = 1000$  samples of length  $T \in \{500, 1000\}$  and apply our procedure to estimate  $u_0$ . We thus obtain  $N = 1000$  estimates of  $u_0$  for each model specification. The results are presented by histograms that show the empirical distribution of the estimates for each specification. In particular, the bars in the plots give the number of simulations (out of a total of 1000) in which a certain value  $\hat{u}_0$  is obtained.

The simulation results for the design with  $\mu_1$  are presented in the upper part of Figure 2, the left-hand panel corresponding to a sample size of  $T = 500$  and the right-hand one to  $T = 1000$ . Since  $\mu_1$  has a jump at  $u_0 = 0.5$ , it deviates from zero very quickly. Our procedure is thus able to localize the point  $u_0$  quite precisely. This becomes visible in the histograms which show that the estimates are not very dispersed but cluster tightly around  $u_0 = 0.5$ . The plots also make visible a slight upward bias which becomes less pronounced when moving to the larger sample size  $T = 1000$ .

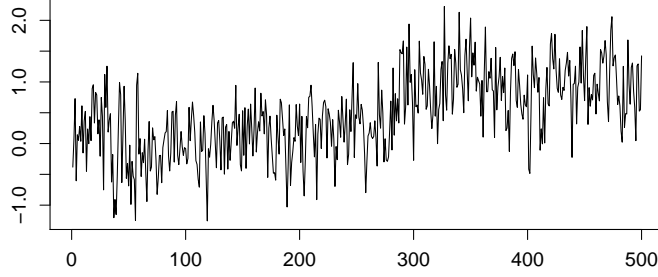


Figure 3: A typical sample path of length  $T = 500$  for model (7.1) with the mean function  $\mu_2$ .

The results for the function  $\mu_2$  are depicted in the lower part of Figure 2. The plots show that the upward bias is more pronounced than in the setting with  $\mu_1$ . This reflects the fact that it is more difficult to localize gradual changes than a jump. In fact, it is quite hard to detect smooth time-variations on the interval  $[0, u_0 + \delta]$  if  $\delta$  is small. This is illustrated by Figure 3, which shows a typical sample path for model (7.1) with mean function  $\mu_2$  of length  $T = 500$ . As can be seen, the deviation of  $\mu_2$  from zero is clearly visible only at time points which are somewhat larger than  $u_0 = 0.5$ . When getting close to  $u_0$ , the signal of time-variation becomes fairly weak and is more and more dominated by the noise of the error term.

In both designs, there is a certain fraction of estimates which take values below  $u_0$ . Theoretically, this fraction should be around 10%, since we have set the probability  $\alpha$  of underestimating  $u_0$  to equal 0.1. In our simulations, however, the fraction obviously lies below the 10% target as can be seen from the plots. This is a small sample effect which can be explained as follows: Our preliminary estimate  $\hat{u}_0^\circ$  is quite conservative, tending to strongly overestimate  $u_0$ . Since  $q_\alpha(u) \geq q_\alpha(u_0)$  for  $u > u_0$ , this implies that the estimate  $\hat{\tau}_\alpha = q_\alpha(\hat{u}_0^\circ)$  will often overshoot the value of the critical threshold  $\tau_\alpha = q_\alpha(u_0)$ , which is used to set up the second step estimator  $\hat{u}_0$ . As a result, the empirical probability of underestimating  $u_0$  tends to lie below the target  $\alpha$  in small samples.

We next investigate the performance of our procedure when the smooth change point  $u_0$  occurs very early in the sample. In particular, we examine the extreme case where  $u_0 = 0$  and the mean function is time-varying over the whole interval  $[0, 1]$ . For this purpose, we consider the setting (7.1) with the mean function

$$\mu_3(u) = 10u \cdot 1(0 \leq u < 0.2) + \{2 - 2.5(u - 0.2)\} \cdot 1(u \geq 0.2). \quad (7.4)$$

The simulation results for this design are depicted in Figure 4 and show that our method detects the time-variation rather quickly. Of course, it is only able to detect it with some delay which becomes smaller when moving to the larger sample size  $T = 1000$ .

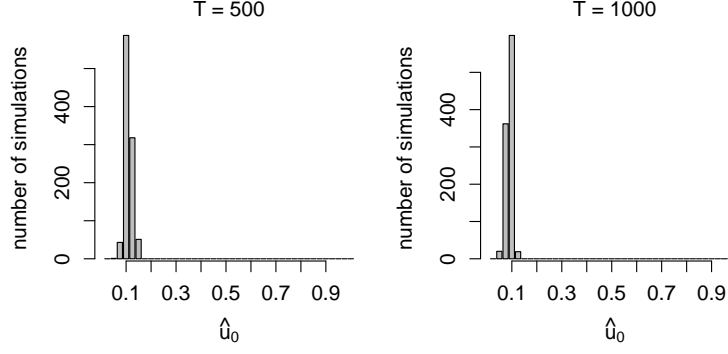


Figure 4: Simulation results produced by our method in model (7.1) with the mean function  $\mu_3$  defined in (7.4).

## 7.2 Comparison with other methods

In this section, we compare our estimation approach with the methods of Mallik et al. (2011, 2013) and Hušková (1999) which are specifically designed to detect gradual changes in the location model (7.1). As before, we assume that the mean function  $\mu$  is constant on the time interval  $[0, u_0]$ , that is,  $\mu(u) = \bar{\mu}$  for  $u \leq u_0$ , and then starts to vary over time. The method of Mallik et al. (2011, 2013) allows to estimate the time point  $u_0$  when  $\mu$  is a smooth nonparametric function that is restricted to take values larger than  $\bar{\mu}$  at time points  $u > u_0$ , that is,  $\mu(u) > \bar{\mu}$  for  $u > u_0$ . The procedure of Hušková (1999) in contrast is based on the parametric assumption that  $\mu(u) = \bar{\mu} + \delta \cdot (u - u_0)^\beta \cdot 1(u > u_0)$  for some slope parameter  $\delta > 0$  and a known constant  $\beta \in [0, 1]$ . In what follows, we set  $\beta = 1$ , thus considering Hušková's method for the class of broken lines with a kink at  $u_0$ .

To compare our method with these two approaches, we set  $u_0 = 0.5$  and consider two different specifications of the mean function  $\mu$ ,

$$\mu_4(u) = 2(u - 0.5) \cdot 1(u > 0.5) \quad (7.5)$$

$$\mu_5(u) = \{10(u - 0.5)\} \cdot 1(0.5 < u < 0.6) + 1(u \geq 0.6). \quad (7.6)$$

Moreover, we let  $\varepsilon_t$  be i.i.d. residuals that are normally distributed with mean zero and standard deviation 0.2. Note that  $\mu_4$  belongs to the parametric family of broken lines for which Hušková's method with  $\beta = 1$  is designed. The function  $\mu_5$ , in contrast, is not an element of this parametric family.

Our estimator is implemented in the same way as in the simulation study of Subsection 7.1. As the error terms are i.i.d., the error variance simplifies to  $\sigma^2 = \mathbb{E}[\varepsilon_t^2]$  and can be estimated as follows: Since  $\mu$  is smooth,  $\mu(\frac{t}{T}) - \mu(\frac{t-1}{T}) = O(T^{-1})$ . This implies that  $X_{t,T} - X_{t-1,T} = \varepsilon_t - \varepsilon_{t-1} + O(T^{-1})$ , which in turn yields that  $\mathbb{E}(X_{t,T} - X_{t-1,T})^2 = \mathbb{E}(\varepsilon_t - \varepsilon_{t-1})^2 + O(T^{-2}) = 2\sigma^2 + O(T^{-2})$ . Hence, we may simply estimate the error variance by  $\hat{\sigma}^2 = T^{-1} \sum_{t=2}^T (X_{t,T} - X_{t-1,T})^2 / 2$ . This estimate is also used in the implementation of the method by Mallik et al. (2011, 2013). Hušková's estimator is constructed as described in equation (1.4) of Hušková (1999). To implement the estimator of Mallik et al. (2011, 2013), we proceed as follows: Since

the method is based on a Nadaraya-Watson smoother of  $\mu$ , we first select the bandwidth  $h$  of this estimator. As shown in Mallik et al. (2013), the rate-optimal bandwidth has the form  $h = cT^{-1/(2k+1)}$ , where  $c$  is a constant and  $\mu$  is assumed to have a cusp of order  $k$  at the point  $u_0$ . This means that the first  $(k-1)$  right derivatives of  $\mu$  at  $u_0$  are zero and the  $k$ -th right derivative is non-zero. For both functions,  $\mu_4$  and  $\mu_5$ ,  $k$  is equal to 1, implying that the optimal bandwidth is of the form  $h = cT^{-1/3}$ . Of course, since the order  $k$  is unknown in practice, this is not a feasible choice of bandwidth. Moreover, even if  $k$  were known, it is not clear how to pick the constant  $c$ . We here ignore these problems and pretend that  $k$  is known. Having repeated the simulations for different choices of the constant  $c$ , we present the results for the choice  $c = 0.1$  which yields the best performance. For simplicity, we also assume that the baseline value  $\bar{\mu}$  is known, so we do not have to replace it by an estimate.

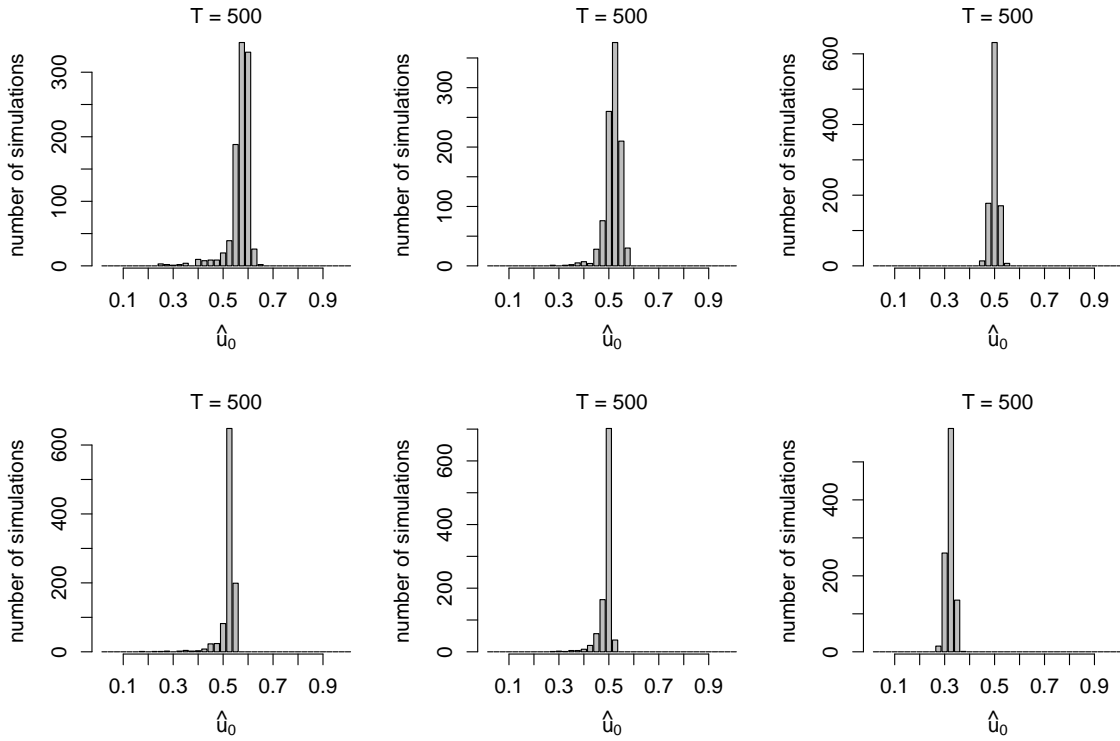


Figure 5: Estimation results for model (7.1) with  $\mu_4$  (upper panel) and  $\mu_5$  (lower panel). The left-hand plots correspond to our method, the middle ones to the approach of Mallik et al. (2011, 2013) and the right-hand ones to the procedure in Hušková (1999).

The results for the regression function  $\mu_4$  are presented in the upper part of Figure 5. As can be seen, Hušková's method outperforms both ours and the  $p$ -value based approach of Mallik et al. (2011, 2013). This is not surprising since it is tailored to a specific parametric class of mean functions to which  $\mu_4$  belongs. Even though less precise than Hušková's estimator, both our and the  $p$ -value based method perform well, ours tending to be a bit more upward biased and thus slightly more conservative. The results for the regression function  $\mu_5$  are presented in the lower part of Figure 5. As before, both our method and that of Mallik et al. perform quite well. The parametric method of Hušková (1999), in contrast, completely

fails to provide reasonable estimates of  $u_0$ . The reason for this is simply that  $\mu_5$  does not satisfy the parametric assumptions of this approach.

To implement the method of Mallik et al. (2011, 2013), we have used an optimally tuned bandwidth which presupposes knowledge of the degree of smoothness  $k$  and have treated the mean value  $\bar{\mu}$  as known. Nevertheless, this approach only provides slightly better results than ours. In practice,  $\bar{\mu}$  must of course be estimated and the optimal choice of bandwidth is not available. Moreover, the performance of the method varies quite considerably with the bandwidth. This is illustrated in Figure 6 which shows the estimation results when picking the bandwidth to be rate optimal with the constants  $c = 0.2, 0.3, 0.4$  instead of  $c = 0.1$ .<sup>1</sup> As can be seen, the results get much worse when slightly changing the bandwidth parameter  $c$ , a large fraction of the estimates tending to strongly underestimate  $u_0$ .

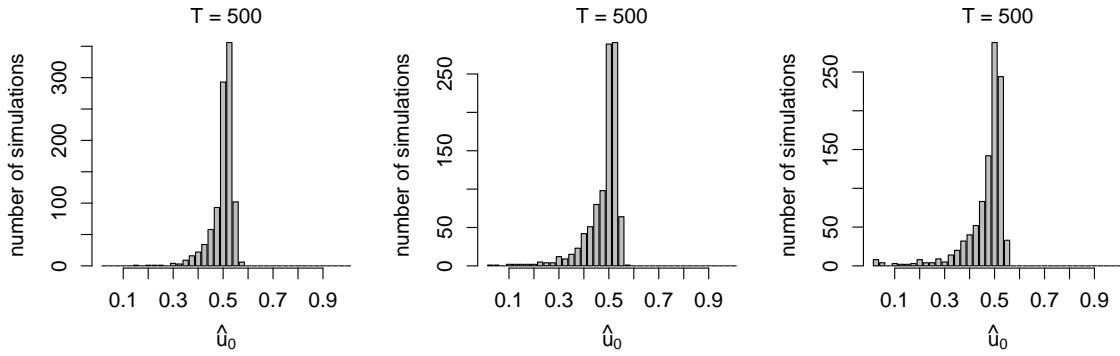


Figure 6: Results for the method of Mallik et al. (2011, 2013) in model (7.1) with  $\mu_4$  and the bandwidth  $h = cT^{-1/3}$ , where  $c = 0.2$  (left),  $c = 0.3$  (middle) and  $c = 0.4$  (right).

The above discussion points to an important advantage of our method: The tuning parameter  $\tau_\alpha$  on which it depends is much more harmless than a bandwidth parameter. As  $\alpha$  can be interpreted in terms of the probability of underestimating  $u_0$ , it is clear how to choose  $\tau_\alpha$  in a reasonable way in practice. Hence, we do not run the risk of producing poor estimates by picking the tuning parameter in an inappropriate way. This makes our procedure particularly attractive to apply in practice. We finally point out that the new method is not specifically designed for detecting a change in the nonparametric location model (7.1) but can be easily adapted to other change point problems. This is illustrated in the Supplementary Material, where we show results for a nonparametric volatility model.

### 7.3 Applications

We now apply the proposed estimation method to the data presented in the Introduction. We first consider the monthly temperature anomalies of the northern hemisphere from 1850 to

<sup>1</sup>Note that for all these values of  $c$ , the bandwidth is fairly small, resulting in an undersmoothed estimate of the mean function  $\mu$ . Specifically, for a sample size of  $T = 500$ , the choice  $c = 0.1$  corresponds to a bandwidth window of approximately 5 data points and  $c = 0.4$  to a window of 25 points. Indeed, the method appears only to work in a reasonable way when strongly undersmoothing, which is already indicated by the fact that the optimal bandwidth is of the rate  $T^{-1/3}$ .

2013 depicted in the left-hand panel of Figure 1. These anomalies are temperature deviations from the average 1961–1990 measured in degrees Celsius. The data set is called HadCRUT4 and can be obtained from the Climatic Research Unit of the University of East Anglia, England. A detailed description of the data can be found in Brohan et al. (2006).

Inspecting the temperature data, they can be seen to exhibit a seasonal as well as a trending behaviour. We thus model them by the equation

$$X_{t,T} = s(t) + \mu\left(\frac{t}{T}\right) + \varepsilon_{t,T} \quad (t = 1, \dots, T), \quad (7.7)$$

where  $s$  is a seasonal component with a period of 12 months,  $\mu$  is a nonparametric trend and  $\varepsilon_{t,T}$  are error terms with zero mean. For identification, we assume that  $\sum_{t=1}^{12} s(t) = 0$ . From the data plot, one can also see a larger variance at the beginning of the sample, suggesting that the errors are nonstationary. To pick up these nonstationary effects, we allow the error terms  $\varepsilon_{t,T}$  to be locally stationary. Rescaling the time argument of the trend component while letting the periodic component depend on real time is a rather natural way to formulate the model. It captures the fact that the trend function is much smoother and varies more slowly than the seasonal part. Analogous model formulations can be found for example in Subba Rao (2004) and Atak et al. (2011).

The issue of global warming has received much attention over the last decades. One question of interest is to locate the onset of the warming trend; see e.g. Thanasis et al. (2011) or Mallik et al. (2011) for a statistical analysis of this question. The challenge is thus to estimate the time point  $u_0$  where the function  $\mu$  starts to strongly trend upwards. To clarify this issue, we apply our estimation method to the anomaly data at hand. Importantly, we do not have to pre-process the data and deseasonalize them but can work with the raw data itself. The reason for this is as follows: The seasonal component  $s$  shows up in averages of the form  $A_T(w) = T^{-1} \sum_{t=1}^{\lfloor wT \rfloor} s(t)$  in the statistic  $\hat{D}_T$  which underlies our estimation procedure. Since  $\sum_{t=1}^{12} s(t) = 0$  by our normalization, it holds that  $A_T(w) = O(T^{-1})$  uniformly in  $w$ . Hence, the seasonal component gets smoothed or averaged out when calculating the statistic  $\hat{D}_T$ , implying that we can simply ignore it.

To implement our procedure, we proceed as described in Section 6 and set  $\alpha = 0.1$ . To calculate the quantiles in Step 1 of the implementation, we use an estimator of the form (6.1) with a bandwidth  $h$  that corresponds to approximately 10 years of data and the bandwidth  $b = 15$ , meaning that we take into account the first 15 autocovariances when computing the HAC estimator. With these choices, we obtain an estimate  $\hat{u}_0$  which corresponds to the year 1915 and is graphically illustrated by the dashed vertical line in the left-hand panel of Figure 1. As a robustness check, we have varied the bandwidth  $h$  between 5 and 15 years and  $b$  between 10 and 20. For all these choices, we obtain estimates roughly between 1910 and 1920, providing evidence that the mean temperature starts to trend upwards in this time region. This finding is in broad accordance with other analyses. Zhao and Woodroffe (2012) for example apply isotonic regression techniques to the data set of yearly global temperature anomalies and find that the warming trend emerges around the same time.

We next turn to the daily return data of the S&P 500 index which are depicted in the right-

hand panel of Figure 1. A simple locally stationary model for financial returns is given by the equation

$$r_{t,T} = \sigma\left(\frac{t}{T}\right)\varepsilon_t, \quad (7.8)$$

where  $r_{t,T}$  denotes the daily return,  $\sigma$  is a time-varying volatility function and  $\varepsilon_t$  are i.i.d. residuals with zero mean and unit variance. Model (7.8) has been studied in a variety of papers [see Drees and Stărică (2003) and Fryzlewicz et al. (2006) among others]. In many situations, it is realistic to assume that the volatility level is more or less constant within some time span  $[u_0, 1]$ , where  $u = 1$  is the present time point, and remains roughly constant in the near future  $(1, 1 + \delta]$ . In this case,  $\sigma(u) \approx \sigma(1)$  at future time points  $u \in (1, 1 + \delta]$ , which suggests to use the present volatility level  $\sigma(1)$  as a forecast for the near future [see Fryzlewicz et al. (2006) among others]. To obtain a good volatility forecast, we thus have to construct a good estimator of  $\sigma(1)$ . If we knew the time point  $u_0$ , we could come up with a very simple and precise estimator. In particular, we could estimate  $\sigma^2(1)$  by the sample variance of the observations contained in the time interval  $[u_0, 1]$ . In practice, however, the time point  $u_0$  is not observed and has to be estimated.

In what follows, we estimate the time span  $[u_0, 1]$  where the volatility level of the S&P 500 returns from Figure 1 is more or less constant. To do so, we have to reformulate our estimation method, since it is designed to apply to time spans of the form  $[0, u_0]$  rather than  $[u_0, 1]$ . Since this is trivial to achieve and simply a matter of notation, we neglect the details. As time-variation in the volatility is equivalent to time-variation in the variance  $\text{Var}(r_{t,T}) = \mathbb{E}[r_{t,T}^2]$ , we set up our procedure to detect changes in the variance and implement it as described in Section 6. As before, we let  $\alpha = 0.1$ . Moreover, we choose  $h = 0.1$ , noting again that the results are very robust to different choices of  $h$ . Finally, we set the bandwidth  $b$  to equal zero, assuming that the return data are independent. Our estimate  $\hat{u}_0$  of the time point  $u_0$  is depicted as the vertical dashed line in the right-hand panel of Figure 1.

**Acknowledgements.** This work has been supported in part by the Collaborative Research Center “Statistical modeling of nonlinear dynamic processes” (SFB 823, Teilprojekt A1, C1) of the German Research Foundation. We would like to thank Rainer Dahlhaus, Wolfgang Polonik and Stanislav Volgushev for helpful discussions and comments on an earlier version of this manuscript. We are also grateful to Alina Dette and Martina Stein, who typed parts of this paper with considerable technical expertise. The constructive comments of an associate editor and two referees on an earlier version of this paper led to a substantial improvement of the manuscript. Parts of this paper were written while the authors were visiting the Isaac Newton Institute, Cambridge, UK, in 2014 (“Inference for change-point and related processes”) and the authors would like to thank the institute for its hospitality.

## Appendix

In this appendix, we prove the main theoretical results of the paper. Throughout the appendix, the symbol  $C$  denotes a generic constant which may take a different value on each

occurrence. Moreover, the expression  $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$  is used to denote the  $L_p$ -norm of a real-valued random variable  $X$ .

## Auxiliary Results

Before we turn to the proofs of the main theorems, we derive some technical lemmas which are needed later on. To formulate them, we introduce some additional notation. To start with, partition the observations  $\{X_{t,T} : t = 1, \dots, T\}$  into blocks of size  $q$ , where the  $r$ -th block spans the observations from time point  $(r-1)q+1$  to  $rq$  and we set  $q = CT^b$  for some small  $b > 0$  (in particular  $b < \frac{1}{4}$ ). Now define

$$W_T(k, k') = \sup_{f \in \mathcal{F}} \left| \sum_{r=k}^{k'} Q_{r,T}(f) \right|$$

along with

$$Q_{r,T}(f) = \frac{1}{\sqrt{(k' - k + 1)q}} \sum_{t=(2r-2)q+1}^{(2r-1)q \wedge T} (f(X_{t,T}) - \mathbb{E}f(X_{t,T})).$$

The terms  $Q_{r,T}(f)$  are scaled sums of the variables  $f(X_{t,T}) - \mathbb{E}f(X_{t,T})$ , the summation running over the observations of the  $(2r-1)$ -th block. The expression  $W_T(k, k')$  sums up the terms  $Q_{k,T}(f), \dots, Q_{k',T}(f)$  which correspond to the odd blocks  $(2k-1), (2k+1), (2k+3), \dots, (2k'-1)$ . The next two lemmas provide a bound on the  $L_p$ -norm of  $W_T(k, k')$ .

**Lemma A.1.** *Let assumptions (C1) and (C2) be satisfied and let  $f_0 \in \mathcal{F}$  have the property that  $\mathbb{E}|f_0(X_{t,T})|^{(1+\delta)p} \leq C$  for some even  $p \in \mathbb{N}$  and a small  $\delta > 0$ . Then*

$$\left\| \sum_{r=k}^{k'} Q_{r,T}(f_0) \right\|_p \leq C$$

for some sufficiently large constant  $C$ .

**Proof.** To shorten notation, write  $w_{t,T} = f_0(X_{t,T}) - \mathbb{E}f_0(X_{t,T})$  and consider the term

$$\begin{aligned} V_T &= V_T(k, k') = \mathbb{E} \left[ \left( \sum_{r=k}^{k'} Q_{r,T}(f_0) \right)^p \right] \\ &\leq \frac{1}{((k' - k + 1)q)^{p/2}} \sum_{r_1, \dots, r_p = k}^{k'} \sum_{t_1=(2r_1-2)q+1}^{(2r_1-1)q \wedge T} \dots \sum_{t_p=(2r_p-2)q+1}^{(2r_p-1)q \wedge T} |\mathbb{E}[w_{t_1,T} \dots w_{t_p,T}]| \\ &\leq \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{\substack{t_1, \dots, t_p = (2k-2)q+1 \\ t_1 \leq \dots \leq t_p}}^{(2k'-1)q \wedge T} |\mathbb{E}[w_{t_1,T} \dots w_{t_p,T}]|. \end{aligned}$$

Let  $(t_1, \dots, t_p)$  be a tuple of ordered indices, that is,  $t_1 \leq \dots \leq t_p$ . We say that the index  $t_i$  has a neighbour if  $|t_i - t_{i-1}| \leq C^* \log T$  or  $|t_i - t_{i+1}| \leq C^* \log T$  for some large constant  $C^*$  to be specified later on. Moreover,  $t_i$  is said to have exactly one neighbour if either



$|t_i - t_{i-1}| \leq C^* \log T$  and  $|t_i - t_{i+1}| > C^* \log T$  or vice versa. Finally, we call  $(t_{i-1}, t_i)$  a pair of neighbours if  $|t_i - t_{i-1}| \leq C^* \log T$ . Now let  $S_{\leq}$  denote the set of ordered tuples  $(t_1, \dots, t_p) \in \{(2k-2)q+1, \dots, (2k'-1)q \wedge T\}^p$  such that each index  $t_i$  has a neighbour. In addition, let  $S_{>}$  be the set of tuples such that at least one index does not have a neighbour. With this notation at hand, we can write  $V_T = V_T^{\leq} + V_T^{>}$ , where for  $\ell \in \{\leq, >\}$ ,

$$V_T^{\ell} = \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{(t_1, \dots, t_p) \in S_{\ell}} |\mathbb{E}[w_{t_1, T} \dots w_{t_p, T}]|.$$

We now analyze the two terms  $V_T^{\leq}$  and  $V_T^{>}$  separately. For the investigation of  $V_T^{\leq}$ , define

$$S_{\leq, a} = \{(t_1, \dots, t_p) \in S_{\leq} \mid \text{each index } t_i \text{ has exactly one neighbour}\}$$

together with  $S_{\leq, b} = S_{\leq} \setminus S_{\leq, a}$ . First suppose that  $(t_1, \dots, t_p) \in S_{\leq, a}$ . In this case, there are exactly  $p/2$  pairs  $(t_{2i-1}, t_{2i})$  of neighbours (recalling that  $p$  is even by assumption). Using Davydov's inequality (see e.g. Corollary 1.1 in Bosq (1996)) to bound the covariances of the mixing variables  $w_{t, T}$ , we obtain that

$$\begin{aligned} |\mathbb{E}[w_{t_1, T} \dots w_{t_p, T}]| &\leq |\mathbb{E}[w_{t_1, T} w_{t_2, T}] \mathbb{E}[w_{t_3, T} \dots w_{t_p, T}]| + |\text{Cov}(w_{t_1, T} w_{t_2, T}, w_{t_3, T} \dots w_{t_p, T})| \\ &= |\mathbb{E}[w_{t_1, T} w_{t_2, T}] \mathbb{E}[w_{t_3, T} \dots w_{t_p, T}]| + O(\alpha(C^* \log T)) \\ &= |\text{Cov}(w_{t_1, T}, w_{t_2, T}) \mathbb{E}[w_{t_3, T} \dots w_{t_p, T}]| + O(\alpha(C^* \log T)) \\ &\vdots \\ &\leq \left| \prod_{i=1}^{p/2} \text{Cov}(w_{t_{2i-1}, T}, w_{t_{2i}, T}) \right| + O(T^{-\nu}), \end{aligned}$$

where we have used the fact that the mixing coefficients are decaying exponentially fast and the constant  $\nu > 0$  can be made arbitrarily large (by choosing the constant  $C^*$  sufficiently large). This implies that

$$\begin{aligned} V_T^{\leq, a} &= \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{(t_1, \dots, t_p) \in S_{\leq, a}} |\mathbb{E}[w_{t_1, T} \dots w_{t_p, T}]| \\ &\leq \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{(t_1, \dots, t_p) \in S_{\leq, a}} \left| \prod_{i=1}^{p/2} \text{Cov}(w_{t_{2i-1}, T}, w_{t_{2i}, T}) \right| + o(1) \\ &\leq \frac{p!}{((k' - k + 1)q)^{p/2}} \prod_{i=1}^{p/2} \left( \sum_{\ell=0}^{\lceil C^* \log T \rceil} \sum_{t_{2i-1}=(2k-2)q+1}^{(2k'-1)q \wedge T} |\text{Cov}(w_{t_{2i-1}, T}, w_{t_{2i-1}+\ell, T})| \right) + o(1) \\ &\leq C \frac{p!}{((k' - k + 1)q)^{p/2}} ((k' - k + 1)q)^{p/2} \left( \sum_{\ell=0}^{\lceil C^* \log T \rceil} \alpha(\ell) \right)^{p/2} + o(1) \leq C \end{aligned}$$

for some sufficiently large constant  $C$ , where the last line again uses Davydov's inequality to bound the covariance expressions in the formula.

Next consider the sum  $V_T^{\leq, b}$  corresponding to indices in the set  $S_{\leq, b}$ . The cardinality of this set is bounded by  $C((k' - k + 1)q)^{\frac{p}{2}-1}(\log T)^{\frac{p}{2}+1}$ , which implies

$$V_T^{\leq, b} = \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{(t_1, \dots, t_p) \in S_{\leq, b}} |\mathbb{E}[w_{t_1, T} \dots w_{t_p, T}]| \leq C \frac{(\log T)^{p/2+1}}{(k' - k + 1)q} = o(1)$$

(noting that  $q = T^b$ ). This shows that the term  $V_T^{\leq}$  is bounded.

Finally, we examine the term  $V_T^>$  corresponding to the index set  $S_>$ . By definition, the tuples contained in this set have at least one element, say  $t_i$ , without a neighbour, that is,  $|t_i - t_{i+1}| > C^* \log T$  and  $|t_i - t_{i-1}| > C^* \log T$ . Exploiting the mixing conditions on the model variables in a similar way as above, we obtain that

$$\begin{aligned} \mathbb{E}[w_{t_1, T} \dots w_{t_p, T}] &= \mathbb{E}[w_{t_1, T} \dots w_{t_{i-1}, T}] \mathbb{E}[w_{t_i, T} \dots w_{t_p, T}] + \text{Cov}(w_{t_1, T} \dots w_{t_{i-1}, T}, w_{t_i, T} \dots w_{t_p, T}) \\ &= \mathbb{E}[w_{t_1, T} \dots w_{t_{i-1}, T}] \text{Cov}(w_{t_i, T}, w_{t_{i+1}, T} \dots w_{t_p, T}) + O(T^{-\nu}) = O(T^{-\nu}), \end{aligned}$$

where  $\nu$  can be chosen arbitrarily large (if  $C^*$  is chosen large enough). Recalling the definition of  $V_T^>$ , this yields that  $V_T^> = o(1)$ . Putting everything together, the quantity  $V_T$  is seen to be bounded. This completes the proof.  $\square$

**Lemma A.2.** *Let (C1) and (C2) be satisfied. Moreover, assume that for some even  $p \in \mathbb{N}$  and some small  $\delta > 0$ ,*

$$\mathbb{E} \left[ \left| \frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f, f')} \right|^{(1+\delta)p} \right] \leq C$$

*for all functions  $f, f' \in \mathcal{F}$ . Then for any  $f_0 \in \mathcal{F}$ ,*

$$\|W_T(k, k')\|_p \leq C \left( \left\| \sum_{r=k}^{k'} Q_{r,T}(f_0) \right\|_p + \int_0^{\text{diam}(\mathcal{F})} \mathcal{N}(w/2, \mathcal{F}, d_{\mathcal{F}})^{1/p} dw \right),$$

*where  $\mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})$  is the covering number of  $(\mathcal{F}, d_{\mathcal{F}})$  and  $\text{diam}(\mathcal{F}) = \sup_{f, f' \in \mathcal{F}} d_{\mathcal{F}}(f, f')$  denotes the diameter of  $\mathcal{F}$ .*

**Proof.** The claim immediately follows from Theorem 2.2.4 and Corollary 2.2.5 in van der Vaart and Wellner (1996) (see their remark on p.100 before Subsection 2.2.1). It thus suffices to verify the conditions of Theorem 2.2.4. In particular, we have to show that

$$\mathbb{E} \left[ \left| \sum_{r=k}^{k'} Q_{r,T}(f) - \sum_{r=k}^{k'} Q_{r,T}(f') \right|^p \right] \leq C d_{\mathcal{F}}(f, f')^p$$

for some sufficiently large constant  $C$ . To prove this, we introduce the notation

$$w_{t,T} = \frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f, f')} - \mathbb{E} \left[ \frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f, f')} \right]$$

and consider

$$\begin{aligned}
V_T &= V_T(k, k') = \mathbb{E} \left[ \left| \sum_{r=k}^{k'} \frac{Q_{r,T}(f) - Q_{r,T}(f')}{d_{\mathcal{F}}(f, f')} \right|^p \right] \\
&\leq \frac{1}{((k' - k + 1)q)^{p/2}} \sum_{r_1, \dots, r_p = k}^{k'} \sum_{t_1 = (2r_1 - 2)q + 1}^{(2r_1 - 1)q \wedge T} \cdots \sum_{t_p = (2r_p - 2)q + 1}^{(2r_p - 1)q \wedge T} |\mathbb{E}[w_{t_1, T} \cdots w_{t_p, T}]| \\
&\leq \frac{p!}{((k' - k + 1)q)^{p/2}} \sum_{\substack{t_1, \dots, t_p = (2k - 2)q + 1 \\ t_1 \leq \dots \leq t_p}}^{(2k' - 1)q \wedge T} |\mathbb{E}[w_{t_1, T} \cdots w_{t_p, T}]|.
\end{aligned}$$

Repeating the arguments from Lemma A.1, we can show that  $V_T$  is bounded, thus completing the proof.  $\square$

## Proof of Theorem 5.1

To show that  $\hat{H}_T = \sqrt{T}[\hat{D}_T - D]$  weakly converges to  $H$ , it suffices to prove that

$$\hat{H}_T^c := \sqrt{T}[\hat{D}_T - \mathbb{E}\hat{D}_T] \rightsquigarrow H \quad (\text{A.1})$$

together with

$$\sqrt{T} \sup_{(u, v, f) \in \Delta \times \mathcal{F}} |\mathbb{E}\hat{D}_T - D| = o(1), \quad (\text{A.2})$$

where  $\hat{H}_T^c$  is the centred version of  $\hat{H}_T$ . We start with the proof of (A.2). Making use of condition (C4), we obtain that

$$\begin{aligned}
\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor uT \rfloor} \mathbb{E}[f(X_{t,T})] &= \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor uT \rfloor} \mathbb{E} \left[ f \left( X_t \left( \frac{t}{T} \right) \right) \right] + o(1) \\
&= \sqrt{T} \sum_{t=1}^{\lfloor uT \rfloor} \int_{\frac{t-1}{T}}^{\frac{t}{T}} \mathbb{E}[f(X_t(w))] dw + o(1) \\
&= \sqrt{T} \int_0^u \mathbb{E}[f(X_t(w))] dw + o(1)
\end{aligned}$$

uniformly with respect to  $u \in [0, 1]$  and  $f \in \mathcal{F}$ . From this, (A.2) immediately follows. To verify (A.1), we show weak convergence of the finite dimensional distributions of  $\hat{H}_T^c$  as well as stochastic equicontinuity of  $\hat{H}_T^c$ . In particular, we derive the following two results.

**Proposition A.1.** *For any finite number of points  $(u_i, v_i, f_i)$  with  $1 \leq i \leq n$ , it holds that*

$$(\hat{H}_T^c(u_1, v_1, f_1), \dots, \hat{H}_T^c(u_n, v_n, f_n))^\top \xrightarrow{d} N(0, \Sigma)$$

where  $\Sigma = (\Sigma_{ij})_{1 \leq i, j \leq n}$  and  $\Sigma_{ij} = \text{Cov}(H(u_i, v_i, f_i), H(u_j, v_j, f_j))$ .

**Proposition A.2.** *The sequence of processes  $\hat{H}_T^c$  is asymptotically stochastically equicontinuous, that is, for any  $\varepsilon > 0$ ,*

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \sup_{\substack{|u-u'|+|v-v'| \\ +d_{\mathcal{F}}(f,f') \leq \delta}} |\hat{H}_T^c(u, v, f) - \hat{H}_T^c(u', v', f')| > \varepsilon \right) = 0.$$

To prove these two results, we make use of the notation

$$\hat{H}_T^c(u, v, f) = \hat{G}_T(v, f) - \left(\frac{v}{u}\right) \hat{G}_T(u, f), \quad (\text{A.3})$$

where

$$\hat{G}_T(u, f) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor uT \rfloor} (f(X_{t,T}) - \mathbb{E}f(X_{t,T})). \quad (\text{A.4})$$

Combining Propositions A.1 and A.2, the statement (A.1) follows from a standard functional central limit theorem (see van der Vaart and Wellner (1996)).

**Proof of Proposition A.1.** The proof proceeds in two steps. In the first, we calculate the asymptotic covariances of the process  $\hat{H}_T^c$ , which is achieved by exploiting the locally stationary structure of the model variables. In the second, we apply a central limit theorem for mixing arrays. The details can be found in the Supplementary Material.  $\square$

**Proof of Proposition A.2.** Straightforward calculations show that

$$\begin{aligned} \sup_{\substack{|u-u'|+|v-v'| \\ +d_{\mathcal{F}}(f,f') \leq \delta}} |\hat{H}_T^c(u, v, f) - \hat{H}_T^c(u', v', f')| &\leq 2 \sup_{\substack{|u-u'| \leq \delta \\ f \in \mathcal{F}}} |\hat{G}_T(u, f) - \hat{G}_T(u', f)| \\ &\quad + 2 \sup_{\substack{d_{\mathcal{F}}(f,f') \leq \delta \\ u \in [0,1]}} |\hat{G}_T(u, f) - \hat{G}_T(u, f')| \\ &\quad + 2 \sup_{\substack{u \in [0,1] \\ f \in \mathcal{F}}} |\delta^{\frac{1}{2}-\eta} \hat{G}_T(u, f)| + 2 \sup_{\substack{u \in [0, \delta^{1/2+\eta}] \\ f \in \mathcal{F}}} |\hat{G}_T(u, f)| \end{aligned}$$

for some small  $\eta > 0$ . Therefore, stochastic equicontinuity follows from the statements

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \sup_{\substack{|u-u'| \leq \delta \\ f \in \mathcal{F}}} |\hat{G}_T(u, f) - \hat{G}_T(u', f)| > \varepsilon \right) = 0 \quad (\text{A.5})$$

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \sup_{\substack{d_{\mathcal{F}}(f,f') \leq \delta \\ u \in [0,1]}} |\hat{G}_T(u, f) - \hat{G}_T(u, f')| > \varepsilon \right) = 0 \quad (\text{A.6})$$

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \sup_{\substack{u \in [0,1] \\ f \in \mathcal{F}}} |\delta^{\frac{1}{2}-\eta} \hat{G}_T(u, f)| > \varepsilon \right) = 0 \quad (\text{A.7})$$

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \sup_{\substack{u \in [0, \delta^{1/2+\eta}] \\ f \in \mathcal{F}}} |\hat{G}_T(u, f)| > \varepsilon \right) = 0. \quad (\text{A.8})$$

(A.5)–(A.8) can be shown by very similar arguments. We thus restrict ourselves to the proof of (A.5).

First of all, observe that for any function  $g : [0, 1] \rightarrow \mathbb{R}$ , the inequality

$$\begin{aligned} \sup_{\substack{|u-u'|\leq\delta \\ u,u'\in[0,1]}} |g(u) - g(u')| &\leq \max_{j=1,\dots,\lceil 1/\delta \rceil} \sup_{u\in[u_{j-1},u_j]} |g(u) - g(u_j)| \\ &\quad + \max_{j=1,\dots,\lceil 1/\delta \rceil} \sup_{u'\in[u_{j-2},u_{j+1}]} |g(u') - g(u_j)| \end{aligned}$$

holds, where  $u_{-1} = u_0 = 0$ ,  $u_j = j\delta$  ( $j = 1, \dots, \lceil 1/\delta \rceil - 1$ ) and  $u_{\lceil 1/\delta \rceil} = u_{\lceil 1/\delta \rceil+1} = 1$ . From this, it is easily seen that (A.5) is a consequence of

$$\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} \mathbb{P} \left( \max_{j=1,\dots,\lceil 1/\delta \rceil} \sup_{u\in[u_{j-1},u_j]} \sup_{f\in\mathcal{F}} \left| \hat{G}_T(u, f) - \hat{G}_T(j\delta, f) \right| > \varepsilon \right) = 0. \quad (\text{A.9})$$

In the sequel, we derive a suitable bound for the probability

$$P_T(\delta, \varepsilon) = \mathbb{P} \left( \max_{j=1,\dots,\lceil 1/\delta \rceil} \sup_{u\in[u_{j-1},u_j]} \sup_{f\in\mathcal{F}} \left| \hat{G}_T(u, f) - \hat{G}_T(j\delta, f) \right| > \varepsilon \right)$$

in (A.9). To start with, we crudely bound this probability by  $P_T(\delta, \varepsilon) \leq \sum_{j=1}^{\lceil 1/\delta \rceil} P_{T,j}(\delta, \varepsilon)$ , where

$$\begin{aligned} P_{T,j}(\delta, \varepsilon) &= \mathbb{P} \left( \sup_{u\in[u_{j-1},u_j]} \sup_{f\in\mathcal{F}} \left| \hat{G}_T(u, f) - \hat{G}_T(j\delta, f) \right| > \varepsilon \right) \\ &= \mathbb{P} \left( \max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \sup_{f\in\mathcal{F}} \left| \hat{G}_T\left(\frac{\ell}{T}, f\right) - \hat{G}_T(j\delta, f) \right| > \varepsilon \right). \end{aligned}$$

To bound the probabilities  $P_{T,j}(\delta, \varepsilon)$ , we write

$$\hat{G}_T(j\delta, f) - \hat{G}_T\left(\frac{\ell}{T}, f\right) = B_T^{\ell+}(f) + \sum_{r=\lceil \frac{\ell}{q} \rceil + 1}^{\lfloor \frac{j\delta T}{q} \rfloor} B_{r,T}(f) + B_T^{j-}(f).$$

Here,  $B_{r,T}(f)$  are blocks of length  $q$  given by

$$B_{r,T}(f) = \frac{1}{\sqrt{T}} \sum_{t=(r-1)q+1}^{rq} (f(X_{t,T}) - \mathbb{E}f(X_{t,T})),$$

where as in the subsection on auxiliary results, we set  $q = CT^b$  for some small  $b > 0$  (specifically,  $b < \frac{1}{4}$ ). In addition,

$$\begin{aligned} B_T^{\ell+}(f) &= \frac{1}{\sqrt{T}} \sum_{t=\ell+1}^{\lceil \frac{\ell}{q} \rceil q} (f(X_{t,T}) - \mathbb{E}f(X_{t,T})) \\ B_T^{j-}(f) &= \frac{1}{\sqrt{T}} \sum_{t=\lfloor \frac{j\delta T}{q} \rfloor q+1}^{\lfloor j\delta T \rfloor} (f(X_{t,T}) - \mathbb{E}f(X_{t,T})) \end{aligned}$$

denote the first and the last block, respectively. With this notation at hand, we obtain

$$\begin{aligned}
P_{T,j}(\delta, 6\varepsilon) &\leq \mathbb{P}\left(\max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \sup_{f \in \mathcal{F}} \left| \sum_{r=\lceil \frac{\ell}{q} \rceil+1}^{\lfloor \frac{j\delta T}{q} \rfloor} B_{r,T}(f) \right| > 4\varepsilon\right) \\
&\quad + \mathbb{P}\left(\max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| > \varepsilon\right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}} |B_T^{j-}(f)| > \varepsilon\right) \\
&=: P_{T,j,1}(\delta, 4\varepsilon) + P_{T,j,2}(\delta, \varepsilon) + P_{T,j,3}(\delta, \varepsilon).
\end{aligned}$$

The terms  $P_{T,j,2}$  and  $P_{T,j,3}$  can be bounded by fairly straightforward arguments: Applying a maximal inequality (see e.g. Section 2.1.3 in van der Vaart and Wellner (1996)), we get that

$$\left\| \max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| \right\|_p \leq C(\delta T)^{1/p} \max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \left\| \sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| \right\|_p.$$

Moreover,

$$\sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| \leq \frac{2}{\sqrt{T}} \sum_{t=\ell+1}^{\lceil \frac{\ell}{q} \rceil q} F(X_{t,T})$$

and by the moment conditions on the envelope  $F$  in (C3),  $\left\| \sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| \right\|_p \leq Cq/\sqrt{T}$ . Hence by Markov's inequality,

$$P_{T,j,2}(\delta, \varepsilon) \leq \varepsilon^{-p} \left\| \max_{\lfloor (j-1)\delta T \rfloor \leq \ell \leq \lfloor j\delta T \rfloor} \sup_{f \in \mathcal{F}} |B_T^{\ell+}(f)| \right\|_p^p \leq C\delta T \left( \frac{q}{\varepsilon\sqrt{T}} \right)^p = o(1)$$

for  $T \rightarrow \infty$  given that  $q = T^b$  with  $b < \frac{1}{4}$ . By similar considerations,  $P_{T,j,3}(\delta, \varepsilon)$  is seen to converge to zero as well. To deal with  $P_{T,j,1}$ , we split it up into two parts:

$$P_{T,j,1}(\delta, 4\varepsilon) \leq \Delta_T^{(0)} + \Delta_T^{(1)}$$

with

$$\begin{aligned}
\Delta_T^{(0)} &= \mathbb{P}\left(\max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \sup_{f \in \mathcal{F}} \left| \sum_{r=k}^{\lfloor \frac{j\delta T}{2q} \rfloor} B_{2r,T}(f) \right| > 2\varepsilon\right) \\
\Delta_T^{(1)} &= \mathbb{P}\left(\max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \sup_{f \in \mathcal{F}} \left| \sum_{r=k}^{\lceil \frac{j\delta T}{2q} \rceil} B_{2r-1,T}(f) \right| > 2\varepsilon\right).
\end{aligned}$$

As the two terms can be treated in the same way, we restrict ourselves to  $\Delta_T^{(1)}$ . Applying a version of Ottaviani's inequality for  $\alpha$ -mixing processes (which has the form stated in Chapter 10.2 of Lin and Bai (2010) and can be proven by the arguments therein), we obtain

that

$$\Delta_T^{(1)} \leq \frac{\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{r=\lfloor \frac{(j-1)\delta T}{2q} \rfloor}^{\lceil \frac{j\delta T}{2q} \rceil} B_{2r-1,T}(f) \right| > \varepsilon\right) + \frac{\delta T}{2q} \alpha(q)}{1 - \max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{r=\lfloor \frac{(j-1)\delta T}{2q} \rfloor}^k B_{2r-1,T}(f) \right| > \varepsilon\right)}. \quad (\text{A.10})$$

In order to bound the right-hand side of (A.10), we make use of the random variables

$$Q_{r,T}(f) = \frac{1}{\sqrt{(k' - k + 1)q}} \sum_{t=(2r-2)q+1}^{(2r-1)q \wedge T} (f(X_{t,T}) - \mathbb{E}f(X_{t,T}))$$

and  $W_T(k, k') = \sup_{f \in \mathcal{F}} |\sum_{r=k}^{k'} Q_{r,T}(f)|$ , which have been introduced at the beginning of the appendix. Combining Lemmas A.1 and A.2 and noting that  $\int_0^{\text{diam}(\mathcal{F})} \mathcal{N}(w/2, \mathcal{F}, d)^{1/p} dw$  is finite by assumption (C3), we get that  $\mathbb{E}[|W_T(k, k')|^p] \leq C < \infty$  for some sufficiently large constant  $C$ . This implies that

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{r=k}^{k'} B_{2r-1,T}(f) \right| > \varepsilon\right) &= \mathbb{P}\left(W_T(k, k') > \frac{\varepsilon \sqrt{T}}{\sqrt{(k' - k + 1)q}}\right) \\ &\leq \mathbb{E}[|W_T(k, k')|^p] \left(\frac{(k' - k + 1)q}{\varepsilon^2 T}\right)^{p/2} \leq C \left(\frac{(k' - k + 1)q}{\varepsilon^2 T}\right)^{p/2}. \end{aligned}$$

Specifically, whenever  $(k - k' + 1)q \leq \delta T$ ,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{r=k}^{k'} B_{2r-1,T}(f) \right| > \varepsilon\right) \leq C \frac{\delta^{p/2}}{\varepsilon^p}. \quad (\text{A.11})$$

With (A.11), it is easy to see that the denominator in (A.10) is bounded away from zero as  $T \rightarrow \infty$  and to infer that

$$\Delta_T^{(1)} \leq C \left( \frac{\delta^{p/2}}{\varepsilon^p} + \frac{\delta T}{2q} \alpha(q) \right).$$

Using an analogous bound for the term  $\Delta_T^{(0)}$ , it follows that

$$P_T(\delta, \varepsilon) \leq \sum_{j=1}^{\lceil 1/\delta \rceil} P_{T,j}(\delta, \varepsilon) \leq C \left\lceil \frac{1}{\delta} \right\rceil \left( \frac{\delta^{p/2}}{\varepsilon^p} + \frac{\delta T}{2q} \alpha(q) \right).$$

This yields that  $\lim_{\delta \searrow 0} \limsup_{T \rightarrow \infty} P_T(\delta, \varepsilon) = 0$  and the assertion (A.9) follows. By the discussion at the beginning of this proof we obtain (A.5), which implies stochastic equicontinuity.  $\square$

### Proof of Theorem 5.3

The proof is an immediate consequence of the following two statements:

$$\mathbb{P}(\hat{u}_0(\tau_T) < u_0) = o(1) \quad (\text{A.12})$$

$$\mathbb{P}(\hat{u}_0(\tau_T) > u_0 + K\gamma_T) = o(1) \quad (\text{A.13})$$

for some sufficiently large constant  $K > 0$ .

**Proof of (A.12).** It holds that

$$\begin{aligned} \mathbb{P}(\hat{u}_0(\tau_T) < u_0) &\leq \mathbb{P}\left(\sqrt{T}\hat{\mathcal{D}}_T(u) > \tau_T \text{ for some } u < u_0\right) \\ &\leq \mathbb{P}\left(\sqrt{T}\mathcal{D}(u) + \hat{\mathcal{H}}_T(u) > \tau_T \text{ for some } u < u_0\right) \leq \mathbb{P}\left(\sup_{u \in [0,1]} \hat{\mathcal{H}}_T(u) > \tau_T\right), \end{aligned}$$

where the second inequality follows from the fact that  $\sqrt{T}\hat{\mathcal{D}}_T(u) \leq \sqrt{T}\mathcal{D}(u) + \hat{\mathcal{H}}_T(u)$  and the third one exploits the fact that  $\mathcal{D}(u) = 0$  at points  $u < u_0$ . From Corollary 5.2, we know that  $\sup_{u \in [0,1]} \hat{\mathcal{H}}_T(u) = \hat{\mathbb{H}}_T(1)$  converges in distribution to  $\mathbb{H}(1)$ . Moreover, the distribution function  $F$  of  $\mathbb{H}(1)$  is continuous on  $[0, \infty)$  by the results of Section 3 in Lifshits (1982). We can thus infer that the distribution function  $F_T$  of  $\hat{\mathbb{H}}_T(1)$  uniformly converges to  $F$  on  $[0, \infty)$ . As a result, we obtain that

$$\mathbb{P}(\hat{\mathbb{H}}_T(1) > \tau_T) = 1 - F_T(\tau_T) = [1 - F(\tau_T)] + [F(\tau_T) - F_T(\tau_T)] = o(1),$$

which in turn yields (A.12).  $\square$

**Proof of (A.13).** Similarly as above, we can write

$$\begin{aligned} \mathbb{P}(\hat{u}_0(\tau_T) > u_0 + K\gamma_T) &\leq \mathbb{P}\left(\sqrt{T}\hat{\mathcal{D}}_T(u) \leq \tau_T \text{ for some } u > u_0 + K\gamma_T\right) \\ &\leq \mathbb{P}\left(\sqrt{T}\mathcal{D}(u) - \hat{\mathcal{H}}_T(u) \leq \tau_T \text{ for some } u > u_0 + K\gamma_T\right), \end{aligned}$$

the last line following from the fact that  $\sqrt{T}\mathcal{D}(u) - \hat{\mathcal{H}}_T(u) \leq \sqrt{T}\hat{\mathcal{D}}_T(u)$ . Next notice that

$$\min_{u \in [u_0 + K\gamma_T, 1]} \mathcal{D}(u) \geq \frac{c_\kappa(K\gamma_T)^\kappa}{2}$$

for sufficiently large  $T$ , which easily follows upon inspection of (5.1). This allows us to infer that

$$\begin{aligned} &\mathbb{P}\left(\sqrt{T}\mathcal{D}(u) - \hat{\mathcal{H}}_T(u) \leq \tau_T \text{ for some } u > u_0 + K\gamma_T\right) \\ &\leq \mathbb{P}\left(\frac{\sqrt{T}c_\kappa(K\gamma_T)^\kappa}{2} - \hat{\mathbb{H}}_T(1) \leq \tau_T\right) \\ &\leq \mathbb{P}\left(\frac{\sqrt{T}c_\kappa(K\gamma_T)^\kappa}{2} - \hat{\mathbb{H}}_T(1) \leq \tau_T, \hat{\mathbb{H}}_T(1) \leq b_T\right) + \mathbb{P}\left(\hat{\mathbb{H}}_T(1) > b_T\right) =: P_1 + P_2, \end{aligned}$$



where  $b_T$  is some diverging sequence of positive numbers satisfying  $b_T/\tau_T \rightarrow 0$ . As already seen in the proof of (A.12), it holds that  $P_2 = o(1)$ . Moreover,  $P_1 = 0$  for sufficiently large  $T$  if we set  $\gamma_T = (\tau_T/\sqrt{T})^{1/\kappa}$  and choose  $K$  to be sufficiently large. This shows (A.13).  $\square$

## Proof of Theorem 5.4 and Corollary 5.5

Due to space constraints, the proofs are deferred to the Supplementary Material.

## References

- ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59** 817–858.
- ANDREWS, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, **61** 128–156.
- ATAK, A., LINTON, O. and XIAO, Z. (2011). A semiparametric panel model for unbalanced data with application to climate change in the united kingdom. *Journal of Econometrics*, **164** 92–115.
- AUE, A., HÖRMANN, S., HORVÁTH, L. and REIMHERR, M. (2009). Break detection in the covariance structure of multivariate time series models. *Annals of Statistics*, **37** 4046–4087.
- AUE, A. and STEINEBACH, J. (2002). A note on estimating the change-point of a gradually changing stochastic process. *Statistics & Probability Letters*, **56** 177–191.
- BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66** 47–78.
- BERKES, I., GOMBAY, E. and HORVÁTH, L. (2009). Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference*, **139** 2044–2063.
- BISSELL, A. F. (1984a). Estimation of linear trend from a cusum chart or tabulation. *Applied Statistics*, **33** 152–157.
- BISSELL, A. F. (1984b). The performance of control charts and cusums under linear trend. *Applied Statistics*, **33** 145–151.
- BLOOMFIELD, P. (1992). Trends in global temperature. *Climatic Change*, **21** 1–16.
- BOSQ, D. (1996). *Nonparametric statistics for stochastic processes*. New York, Springer.
- BROHAN, P., KENNEDY, J. J., HARRIS, I., TETT, S. F. B. and JONES, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, **111**.
- BROWN, R., DURBIN, J. and EVANS, J. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society Series B*, **37** 149–163.
- CHEN, Y., HÄRDLE, W. K. and PIGORSCH, U. (2010). Localized realized volatility modeling. *Journal of the American Statistical Association*, **105** 1376–1393.
- CHOW, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28** 591–605.
- DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics*, **25** 1–37.
- DAHLHAUS, R. and SUBBA RAO, S. (2006). Statistical inference for time-varying ARCH processes. *Annals of Statistics*, **34** 1075–1114.

- DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101** 223–239.
- DE JONG, R. M. and DAVIDSON, J. (2000). Consistency of kernel estimator of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, **68** 407–423.
- DREES, H. and STĂRICĂ, C. (2003). A simple non-stationary model for stock returns. *Preprint*.
- FRYZLEWICZ, P., SAPATINAS, T. and SUBBA RAO, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, **93** 687–704.
- FRYZLEWICZ, P. and SUBBA RAO, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, **17** 320–346.
- GAN, F. F. (1991). Ewma control chart under linear drift. *Journal of Statistical Computation and Simulation*, **38** 181–200.
- GAN, F. F. (1992). Cusum control chart under linear drift. *Journal of the Royal Statistical Society D*, **41** 71–84.
- GOLDENSHLUGER, A., TSYBAKOV, A. and ZEEVI, A. (2006). Optimal change-point estimation from indirect observations. *Annals of Statistics*, **34** 350–372.
- HANSEN, J., RUEDY, R., SATO, M. and LO, K. (2002). Global warming continues. *Science*, **295** 275.
- HINKLEY, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57** 1–17.
- HORVÁTH, L., KOKOSZKA, P. and STEINEBACH, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, **68** 96–119.
- HUŠKOVÁ, M. (1999). Gradual changes versus abrupt changes. *Journal of Statistical Planning and Inference*, **76** 109–125.
- HUŠKOVÁ, M. and STEINEBACH, J. (2002). Asymptotic tests for gradual changes. *Statistics & Decisions*, **20** 137–151.
- JANDHYALA, V., FOTOPOULOS, S., MACNEILL, I. and LIU, P. (2014). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, **34** 423–446.
- KOO, B. and LINTON, O. (2012). Estimation of semiparametric locally stationary diffusion models. *Journal of Econometrics*, **170** 210–233.
- KRÄMER, W., PLOBERGER, W. and ALT, R. (1988). Testing for structural change in dynamic models. *Econometrica*, **56** 1355–1369.
- LIFSHTIS, M. A. (1982). On the absolute continuity of distributions of functionals of random processes. *Theory of Probability & Its Applications*, **27** 600–607.
- LIN, Z. and BAI, Z. (2010). *Probability inequalities*. New York, Springer.
- MALLIK, A., BANERJEE, M. and SEN, B. (2013). Asymptotics for  $p$ -value based threshold estimation in regression settings. *Preprint*.
- MALLIK, A., SEN, B., BANERJEE, M. and MICHAILIDIS, G. (2011). Threshold estimation based on a  $p$ -value framework in dose-response and regression settings. *Biometrika*, **98** 887–900.
- MERCURIO, D. and SPOKOINY, V. (2004). Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics*, **32** 577–602.
- MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Annals of Statistics*, **20** 737–761.

- PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41** 100–115.
- PAGE, E. S. (1955). Control charts with warning lines. *Biometrika*, **42** 243–257.
- RAIMONDO, M. (1998). Minimax estimation of sharp change points. *Annals of Statistics*, **26** 1379–1397.
- SIEGMUND, D. O. and ZHANG, H. (1994). *Confidence regions in broken line regression*, vol. 23 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 292–316.
- SUBBA RAO, S. (2004). On multiple regression models with nonstationary correlated errors. *Biometrika*, **91** 645–659.
- SUBBA RAO, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Advances in Applied Probability*, **38** 1155–1172.
- THANASIS, V., EFTHIMIA, B.-S. and DIMITRIS, K. (2011). Estimation of linear trend onset in time series. *Simulation Modelling Practice and Theory*, **19** 1384–1398.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes*. New York, Springer.
- VOGT, M. (2012). Nonparametric regression for locally stationary time series. *Annals of Statistics*, **40** 2601–2633.
- WIED, D., KRÄMER, W. and DEHLING, H. (2012). Testing for a change in correlation at an unknown point in time using an extended functional delta method. *Econometric Theory*, **28** 570–589.
- ZHAO, O. and WOODROOFE, M. (2012). Estimating a monotone trend. *Statistica Sinica*, **22** 359–378.

# Supplementary Material for “Detecting Gradual Changes in Locally Stationary Processes”

Michael Vogt                      Holger Dette  
University of Konstanz          Ruhr-Universität Bochum

March 18, 2014

## Abstract

In this supplement, we examine the finite sample performance of our method by further simulations. In addition, we provide the technical details and proofs that are omitted in the paper.

## 1 Simulations

In what follows, we continue the simulation study from Section 7.1 of the paper. As announced there, we examine a volatility model together with a multivariate extension of it. The univariate model is

$$X_{t,T} = \sigma\left(\frac{t}{T}\right)\varepsilon_t, \quad (\text{S.1})$$

where  $\sigma$  is a time-varying volatility function and  $\varepsilon_t$  are i.i.d. residuals that are normally distributed with zero mean and unit variance. This is the same model as discussed in the application on the S&P 500 returns in Section 7.3 of the paper. Our aim is to estimate the time point where the volatility function  $\sigma$  starts to vary over time. We consider two different specifications of  $\sigma$ ,

$$\begin{aligned} \sigma_1(u) &= 1(u < 0.5) + 2 \cdot 1(u \geq 0.5) \\ \sigma_2(u) &= 1(u < 0.5) + \{1 + 10(u - 0.5)\} \cdot 1(0.5 < u < 0.6) + 2 \cdot 1(u \geq 0.6), \end{aligned}$$

both of which are equal to 1 on the interval  $[0, 0.5]$  and then start to vary over time. Thus,  $u_0 = 0.5$  in both cases. Analogously to the time-varying mean setting,  $\sigma_1$  has a jump at  $u_0 = 0.5$ , whereas  $\sigma_2$  smoothly deviates from its baseline value 1.

The multivariate extension of model (S.1) is given by the equation

$$X_{t,T} = \Sigma\left(\frac{t}{T}\right)\varepsilon_t, \quad (\text{S.2})$$

where  $X_{t,T} = (X_{t,T,1}, X_{t,T,2})^\top$  are bivariate random variables,  $\Sigma(u)$  is a  $2 \times 2$ -matrix for each time point  $u$  and  $\varepsilon_t = (\varepsilon_{t,1}, \varepsilon_{t,2})^\top$  are bivariate standard normal i.i.d. residuals. Since

$\Sigma^2(\frac{t}{T}) := \Sigma(\frac{t}{T})\Sigma^\top(\frac{t}{T}) = \mathbb{E}[X_{t,T}X_{t,T}^\top]$ , the time-varying matrix  $\Sigma^2(\frac{t}{T})$  is the covariance matrix of  $X_{t,T}$ . Our aim is to estimate the time point where this matrix starts to vary over time. Put differently, we want to localize the time point where the covariance structure of  $X_{t,T}$  starts to change. The stochastic feature of interest is thus the vector of covariances  $\lambda_{t,T} = (\nu_{t,T}^{(1,1)}, \nu_{t,T}^{(1,2)}, \nu_{t,T}^{(2,2)})^\top$ , where  $\nu_{t,T}^{(i,j)} = \mathbb{E}[X_{t,T,i}X_{t,T,j}]$ . We consider two different specifications of the volatility matrix  $\Sigma$ ,

$$\begin{aligned}\Sigma_1(u) &= \sigma_1(u) \cdot A \\ \Sigma_2(u) &= \sigma_2(u) \cdot A,\end{aligned}$$

where

$$AA^\top = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \text{or put differently,} \quad A \approx \begin{pmatrix} 0.87 & -0.5 \\ 0.87 & 0.5 \end{pmatrix}$$

and  $\sigma_1(u)$  along with  $\sigma_2(u)$  are defined above. Both matrices  $\Sigma_1(u)$  and  $\Sigma_2(u)$  are constant on the interval  $[0, 0.5]$  and then start to vary over time. Hence, as in the univariate case,  $u_0 = 0.5$ .

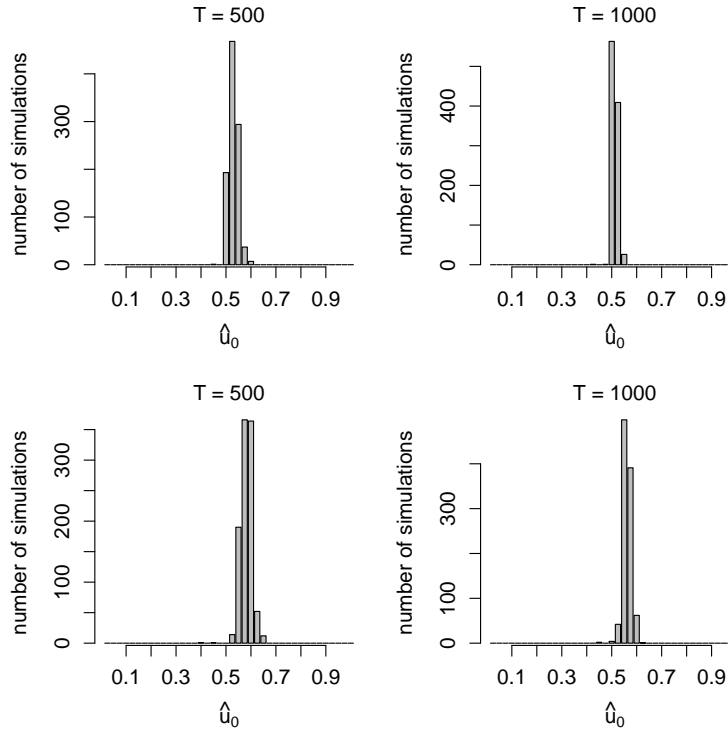


Figure 1: Simulation results for model (S.1) with the volatility function  $\sigma_1$  (upper panel) and the function  $\sigma_2$  (lower panel).

We implement our method as described in Section 6 of the paper, setting the parameter  $\alpha$  to equal 0.1. To calculate the quantiles in the first step of the implementation, we employ an estimator of the form (6.1) with  $h = 0.2$  and a bandwidth  $b$  of zero, exploiting the fact that the simulated data are independent. The resulting estimator is denoted by  $\hat{u}_0$ . For each

model specification, we draw  $N = 1000$  samples of length  $T \in \{500, 1000\}$  and compute the estimate of  $u_0$  for each draw. The results are presented by means of histograms in the same way as in Section 7 of the paper.

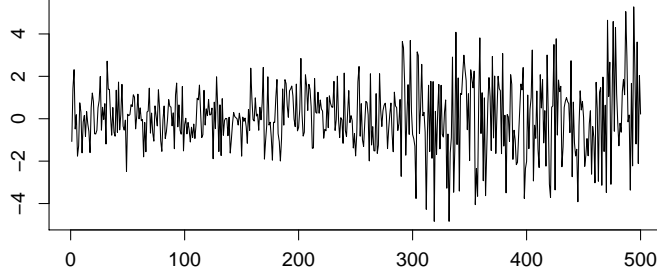


Figure 2: A typical sample path of length  $T = 500$  for model (S.1) with  $\sigma_2$ .

We first discuss the results on the univariate model (S.1). The upper panel of Figure 1 presents the histograms for the design with  $\sigma_1$ , the lower panel those for the design with  $\sigma_2$ . The results are fairly similar to those from the time-varying mean setting: Our method is again able to detect the point  $u_0$  quite precisely in the jump design with  $\sigma_1$ . The histograms in the setup with  $\sigma_2$  are a bit more dispersed, reflecting the fact that it is harder to localize a gradual change than a jump.

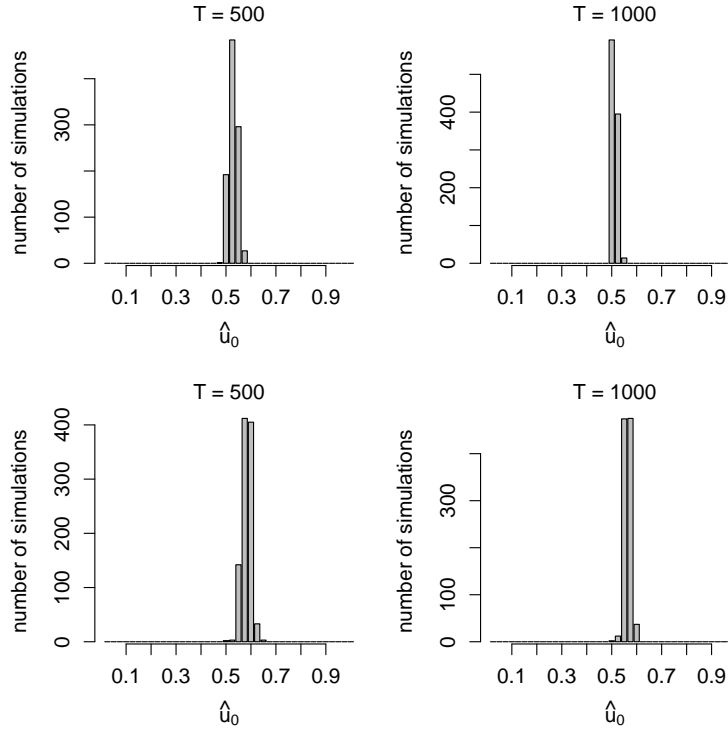


Figure 3: Simulation results for model (S.2) with the volatility matrix  $\Sigma_1$  (upper panel) and the matrix  $\Sigma_2$  (lower panel).

Figure 2 shows a typical sample path of length  $T = 500$  for the design with  $\sigma_2$ . As can be

seen, the increase in the volatility level is hardly visible close to  $u_0 = 0.5$  and only becomes apparent with some delay. It is thus natural that our procedure detects the time-variation in the volatility level only with a bit of delay. This produces the upward bias in the histograms which becomes less pronounced in larger samples.

We finally turn to the results for the bivariate model (S.2). The histograms for the model with  $\Sigma_1$  are displayed in the upper panel of Figure 3, those for the design with  $\Sigma_2$  in the lower panel. Overall, the estimates give a good approximation to the true value  $u_0$ , those in the jump design with  $\Sigma_1$  being a bit more precise than those in the gradual change design. Moreover, the histograms again make visible an upward bias which is comparable in size to that in the univariate setting.

## 2 Technical Details

**Proof of Proposition A.1.** We first calculate the asymptotic expectation and covariances of the process  $\hat{H}_T^c$ . As the process is centered, it holds that  $\mathbb{E}[\hat{H}_T^c(u, v, f)] = 0$ . Moreover,

$$\begin{aligned} \text{Cov}(\hat{H}_T^c(u_1, v_1, f_1), \hat{H}_T^c(u_2, v_2, f_2)) &= \frac{v_1 v_2}{u_1 u_2} \mathbb{E}[\hat{G}_T(u_1, f_1) \hat{G}_T(u_2, f_2)] \\ &\quad - \frac{v_2}{u_2} \mathbb{E}[\hat{G}_T(v_1, f_1) \hat{G}_T(u_2, f_2)] \\ &\quad - \frac{v_1}{u_1} \mathbb{E}[\hat{G}_T(u_1, f_1) \hat{G}_T(v_2, f_2)] \\ &\quad + \mathbb{E}[\hat{G}_T(v_1, f_1) \hat{G}_T(v_2, f_2)]. \end{aligned} \quad (\text{S.3})$$

In what follows, we show that

$$\mathbb{E}[\hat{G}_T(u_1, f_1) \hat{G}_T(u_2, f_2)] = \sum_{\ell=-\infty}^{\infty} \int_0^{\min\{u_1, u_2\}} c_\ell(w) dw + o(1) \quad (\text{S.4})$$

with  $c_\ell(w) = c_\ell(w, f_1, f_2) = \text{Cov}(f_1(X_0(w)), f_2(X_\ell(w)))$ . Plugging (S.4) into (S.3) yields

$$\text{Cov}(\hat{H}_T^c(u_1, v_1, f_1), \hat{H}_T^c(u_2, v_2, f_2)) = \text{Cov}(H(u_1, v_1, f_1), H(u_2, v_2, f_2)) + o(1).$$

Hence, the covariances of  $\hat{H}_T^c$  converge to those of the Gaussian process  $H$ .

To show (S.4), we assume without loss of generality that  $u_1 \leq u_2$ . Exploiting the mixing condition (C2) by means of Davydov's inequality, it can be seen that  $\text{Cov}(f_1(X_{t,T}), f_2(X_{s,T})) \leq C\alpha(|s-t|) \leq Ca^{|s-t|}$  for some  $a < 1$  and a sufficiently large constant  $C$ . We thus obtain that

$$\begin{aligned} &\mathbb{E}[\hat{G}_T(u_1, f_1) \hat{G}_T(u_2, f_2)] \\ &= \frac{1}{T} \sum_{t=1}^{\lfloor u_1 T \rfloor} \sum_{s=1}^{\lfloor u_2 T \rfloor} \text{Cov}(f_1(X_{t,T}), f_2(X_{s,T})) \\ &= \frac{1}{T} \sum_{t=1}^{\lfloor u_1 T \rfloor} \sum_{s=1}^{\lfloor u_2 T \rfloor} I\{|s-t| \leq C^* \log T\} \text{Cov}(f_1(X_{t,T}), f_2(X_{s,T})) + o(1) \\ &=: Q_T^{(1)} + Q_T^{(2)} + Q_T^{(3)} + o(1) \end{aligned}$$

for some sufficiently large constant  $C^*$ , where the random variables  $Q_T^{(j)}$  ( $j = 1, 2, 3$ ) are defined by

$$\begin{aligned} Q_T^{(1)} &= \frac{1}{T} \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} I\{t \leq \lfloor u_1 T \rfloor, t + \ell \leq \lfloor u_2 T \rfloor\} \text{Cov}(f_1(X_{t,T}), f_2(X_{t+\ell,T})) \\ Q_T^{(2)} &= \frac{1}{T} \sum_{t=1}^{\lfloor u_1 T \rfloor} \text{Cov}(f_1(X_{t,T}), f_2(X_{t,T})) \\ Q_T^{(3)} &= \frac{1}{T} \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=\ell+1}^T I\{t \leq \lfloor u_1 T \rfloor, t - \ell \leq \lfloor u_2 T \rfloor\} \text{Cov}(f_1(X_{t,T}), f_2(X_{t-\ell,T})). \end{aligned}$$

By assumption (C4), it follows for  $\ell \leq \lceil C^* \log T \rceil$  and any  $w$  with  $|w - \frac{t}{T}| \leq \frac{1}{T}$  that

$$\begin{aligned} c_{t,T,\ell} &:= \text{Cov}(f_1(X_{t,T}), f_2(X_{t+\ell,T})) \\ &= \text{Cov}\left(f_1\left(X_t\left(\frac{t}{T}\right)\right), f_2\left(X_{t+\ell}\left(\frac{t+\ell}{T}\right)\right)\right) + O\left(\frac{\log T}{T}\right) \\ &= \text{Cov}\left(f_1\left(X_t\left(\frac{t}{T}\right)\right), f_2\left(X_{t+\ell}\left(\frac{t}{T}\right)\right)\right) + O\left(\frac{\log T}{T}\right) \\ &= \text{Cov}(f_1(X_0(w)), f_2(X_\ell(w))) + O\left(\frac{\log T}{T}\right) \\ &=: c_\ell(w) + O\left(\frac{\log T}{T}\right), \end{aligned}$$

the last line defining  $c_\ell(w)$  in an obvious manner. From this, it is easy to see that

$$\begin{aligned} \frac{1}{T} \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} |c_{t,T,\ell}| &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} \left|c_\ell\left(\frac{t}{T}\right)\right| dw + O\left(\frac{(\log T)^2}{T}\right) \\ &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} |c_\ell(w)| dw + O\left(\frac{(\log T)^2}{T}\right) \\ &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \int_0^1 |c_\ell(w)| dw + O\left(\frac{(\log T)^2}{T}\right). \end{aligned}$$

Because of the mixing assumption (C2), the left-hand side of this equation is bounded as  $T \rightarrow \infty$  and consequently  $\sum_{\ell=1}^{\infty} \int_0^1 c_\ell(w) dw$  is absolutely convergent. Therefore we obtain for the term  $Q_T^{(1)}$  as  $T \rightarrow \infty$  (recall that  $u_1 \leq u_2$ )

$$\begin{aligned} Q_T^{(1)} &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{\lfloor u_1 T \rfloor - \ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} c_\ell(w) dw + O\left(\frac{(\log T)^2}{T}\right) \\ &= \sum_{\ell=1}^{\infty} \int_0^{u_1} c_\ell(w) dw + O\left(\frac{(\log T)^2}{T}\right) \end{aligned}$$

and similarly

$$Q_T^{(2)} = \int_0^{u_1} c_0(w) dw + O\left(\frac{(\log T)^2}{T}\right), \quad Q_T^{(3)} = \sum_{\ell=1}^{\infty} \int_0^{u_1} c_{-\ell}(w) dw + O\left(\frac{(\log T)^2}{T}\right).$$



Putting everything together, we arrive at (S.4).

Having calculated the asymptotic covariance structure of  $\hat{H}_T^c$ , we now apply a central limit theorem for mixing arrays of random variables (see e.g. Liebscher (1996)) together with the Cramér-Wold device to obtain weak convergence of the finite dimensional distributions.  $\square$

**Proof of Theorem 5.4.** We first derive (5.7) which says that

$$\mathbb{P}(\hat{u}_0(\tau_\alpha) < u_0) \leq \alpha + o(1).$$

It holds that

$$\begin{aligned} \mathbb{P}(\hat{u}_0(\tau_\alpha) < u_0) &\leq \mathbb{P}(\sqrt{T}\hat{\mathcal{D}}_T(u) > \tau_\alpha \text{ for some } u < u_0) \\ &\leq \mathbb{P}(\sqrt{T}\hat{\mathbb{D}}_T(u_0) > \tau_\alpha) \\ &= \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \tau_\alpha). \end{aligned}$$

We now make use of the following fact which is a direct consequence of the results from Section 3 in Lifshits (1982):

(\*) For each  $u$ , the random variable

$$\mathbb{H}(u) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u} |H(v, w, f)|$$

has a distribution function which is continuous on  $[0, \infty)$ .

By (\*), we obtain that

$$\begin{aligned} \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \tau_\alpha) &= \mathbb{P}(\mathbb{H}(u_0) > \tau_\alpha) + \left[ \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \tau_\alpha) - \mathbb{P}(\mathbb{H}(u_0) > \tau_\alpha) \right] \\ &= \mathbb{P}(\mathbb{H}(u_0) > \tau_\alpha) + o(1) = \alpha + o(1), \end{aligned}$$

where the last equality is due to the fact that  $\tau_\alpha = q_\alpha(u_0)$  is the  $(1 - \alpha)$ -quantile of  $\mathbb{H}(u_0)$ . From this, (5.7) immediately follows. The statement (5.8) can be proven by the same arguments as for (A.13) in the proof of Theorem 5.3.  $\square$

**Proof of Corollary 5.5.** Let  $q_\alpha(u_n)$  be the  $(1 - \alpha)$ -quantile of  $\mathbb{H}(u_n)$  and  $q_\alpha(u)$  the corresponding quantile of  $\mathbb{H}(u)$ . We first show that for any  $\alpha > 0$ ,

$$q_\alpha(u_n) \rightarrow q_\alpha(u) \tag{S.5}$$

as  $u_n \rightarrow u$ . To do so, let  $\mathcal{C}_u(\Delta, d)$  denote the space of uniformly continuous functions on  $(\Delta, d)$  and define the functionals

$$\begin{aligned} M_n(x) &= M_{u_n}(x) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u_n} |x(v, w, f)| \\ M(x) &= M_u(x) = \sup_{f \in \mathcal{F}} \sup_{0 \leq w \leq v \leq u} |x(v, w, f)| \end{aligned}$$

for  $x \in \mathcal{C}_u(\Delta, d)$ . Elementary arguments show that

$$M(x) = \lim_{n \rightarrow \infty, y \rightarrow x} M_n(y),$$

where  $x$  and  $y$  are elements of  $\mathcal{C}_u(\Delta, d)$ . Using this together with the extended continuous mapping theorem (see e.g. Theorem 1.11.1 in van der Vaart and Wellner (1996)), we obtain that

$$M_n(H) \xrightarrow{d} M(H).$$

Noting that  $M_n(H) = \mathbb{H}(u_n)$  and  $M(H) = \mathbb{H}(u)$ , this can be re-expresses as

$$\mathbb{H}(u_n) \xrightarrow{d} \mathbb{H}(u).$$

As the distribution function of  $\mathbb{H}(u)$  is continuous on  $[0, \infty)$  by  $(*)$ , we can conclude that the quantile functions converge as well, thus arriving at (S.5).

Next let  $\tilde{u}_0$  be a consistent estimator of  $u_0$ . By (S.5), the quantile function  $q_\alpha(\cdot)$  is continuous at each point  $u$ , in particular at  $u_0$ . Hence,

$$\hat{\tau}_\alpha = q_\alpha(\tilde{u}_0) \xrightarrow{P} \tau_\alpha = q_\alpha(u_0). \quad (\text{S.6})$$

Moreover,

$$\begin{aligned} \mathbb{P}(\hat{u}_0(\hat{\tau}_\alpha) < u_0) &\leq \mathbb{P}(\sqrt{T}\hat{\mathcal{D}}_T(u) > \hat{\tau}_\alpha \text{ for some } u < u_0) \\ &\leq \mathbb{P}(\sqrt{T}\hat{\mathbb{D}}_T(u_0) > \hat{\tau}_\alpha) \\ &= \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \hat{\tau}_\alpha). \end{aligned}$$

Since  $\hat{\mathbb{H}}_T(u_0) \xrightarrow{d} \mathbb{H}(u_0)$  and the distribution function of  $\mathbb{H}(u_0)$  is continuous on  $[0, \infty)$  by  $(*)$ , the distribution function of  $\hat{\mathbb{H}}_T(u)$  uniformly converges to that of  $\mathbb{H}(u)$  on  $[0, \infty)$ . Hence,

$$\begin{aligned} \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \hat{\tau}_\alpha) &= \mathbb{P}(\mathbb{H}(u_0) > \hat{\tau}_\alpha) + \left[ \mathbb{P}(\hat{\mathbb{H}}_T(u_0) > \hat{\tau}_\alpha) - \mathbb{P}(\mathbb{H}(u_0) > \hat{\tau}_\alpha) \right] \\ &= \mathbb{P}(\mathbb{H}(u_0) > \hat{\tau}_\alpha) + o_p(1). \end{aligned}$$

Finally, as  $\hat{\tau}_\alpha = \tau_\alpha + o_p(1)$  and the distribution function of  $\mathbb{H}(u_0)$  is continuous by  $(*)$ , we obtain that

$$\mathbb{P}(\mathbb{H}(u_0) > \hat{\tau}_\alpha) = \mathbb{P}(\mathbb{H}(u_0) > \tau_\alpha) + o(1) = \alpha + o(1).$$

This completes the proof of (5.9). The statement (5.10) can again be shown by the same arguments as for (A.13) in the proof of Theorem 5.3.  $\square$

## References

- LIEBSCHER, E. (1996). Central limit theorems for sums of  $\alpha$ -mixing random variables. *Stochastics and Stochastic Reports*, **59** 241–258.
- LIFSHITS, M. A. (1982). On the absolute continuity of distributions of functionals of random processes. *Theory of Probability & Its Applications*, **27** 600–607.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes*. New York, Springer.